The Developmental Landscape of Neurogenesis in *Drosophila melanogaster* Revealed Using Targeted Single-Cell RNA-Sequencing and a Multi-Informatic Analysis Paradigm

by

Nigel S. Michki

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biophysics)
in the University of Michigan
2021

Doctoral Committee:

Associate Professor Dawen Cai, chair
Associate Professor Catherine Collins
Associate Professor Cheng-Yu Lee
Associate Professor Kevin Wood
Assistant Professor Qiong Yang

Nigel S. Michki
nigelmic@umich.edu
ORCID iD: 0000-0003-0403-0648

Dedication

For Alyse.

Acknowledgements

I thank Dawen Cai and all of my committee members for their guidance and mentorship over the past 5 years. They have helped mold me from a budding physicist into a blossoming bio/neuro/chemist/informatician/..., and I will forever be grateful to them for making sure I have all of the resources needed in order for me to pursue my goals.

I thank the members of the Cai lab, past and present, all of whom been integral to the development of the tools, techniques, and ideas that I have used and explored over the course of my thesis work. In particular, I thank Daniel Núñez and Douglas Roossien, my postdoctoral mentors, and Ye Li, my graduate student and now postdoctoral colleague, for developing the groundwork of this thesis with me and teaching me many, if not all, of the hands-on skills I use in the lab. I also thank Logan Walker for being the best lab friend I could have asked for, supplying me with the memes and encouragement I needed in order to bring this work to fruition.

I thank my parents, Renee Boyer-Michki and Kevin Michki, for loving me and always encouraging me to do my best. This thesis is as much a product of their hard work in raising me as it is mine in pursuing it, and I can never thank them enough for all they have done for me. I also thank my siblings, Elliott, Heather, and Ryan, for putting up with my antics and for bringing beautiful new music, people, and hairstyles into the world. Finally, I thank Judi Boyer, Tom Boyer, and Heidi Boyer for being a constant source of love, encouragement, and chocolates.

The work described in Chapter 2 is the product of years of collaborative research. To that end, I credit Ye Li, Kayvon Sanjasaz, Yimeng Zhao, Fred Y. Shen, Logan A.

Walker, Wenjia Cao, Cheng-Yu Lee, and Dawen Cai for their experimental, intellectual, and oversight contributions. We (the collective authors of the work referenced therein) acknowledge support from the University of Michigan Flow Cytometry Core and its National Institutes of Health (NIH) support, NIH 5P30CA046592-31, as well as the University of Michigan Advanced Genomics Core. I acknowledge support from NIH 1T32EB005582. F.Y.S. acknowledges the support from NIH 1F31NS11184701. D.C. acknowledges support from the University of Michigan (CDB IDEA Awards in Stem Cell Biology, MCubed2.0) and Michigan Economic Development Corporation (Mi-TRAC). C.-Y.L. acknowledges support from the NIH 1R01NS107496.

Many figures in this work were developed using biorender.com, printed here with permission under a paid academic use license. Additionally, most computational analysis in this work was performed using open source software, credited in large part in appendix A5. I thank the artists and software developers who made conducting and describing this work possible.

Preface

This dissertation is organized into three chapters, roughly encompassing a literature review of neurogenesis mechanisms in *Drosophila melanogaster* (and other model organisms), our work to characterize neural fate patterning in the type II neuroblast lineages of *Drosophila*, and an overview of the state of the field in single-cell -omics experimental design and analysis, the latter of which I expand upon greatly. As such, it covers a range of topics in the rapidly developing single-cell '-omics' field, and I aim to be up-to-date in my descriptions therein at the time of writing. However, even as this field has drastically changed even over the course of my own PhD research, shifts in best practices for experimental design and analysis should be expected in the future as new technologies are pioneered to uncover what makes each and every cell in the world unique.

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

The *Drosophila* type-II neuroblast lineages present an attractive model to investigate the neurogenesis and differentiation process as they adapt to a process similar to that in the human outer subventricular zone. Here, I performed targeted single-cell mRNA sequencing in third instar larval brains to study this process of the type-II neuroblast lineages. Combining prior knowledge, in silico analyses, and in situ validation, my multi-informatic investigation describes the molecular landscape from a single developmental snapshot. 17 markers are identified to differentiate distinct maturation stages. 30 markers are identified to specify the stem cell origin and/or cell division numbers of INPs, and at least 12 neuronal subtypes are identified. To foster future discoveries, I developed MiCV, a web tool for rapidly and interactively analyzing scRNA-seq datasets. Taken together, these resources majorly advance our understanding of the neural differentiation process at the molecular level.

Chapter 1

Introduction: The Evolutionary Scaling of Neurogenesis During Development

Model organisms and their brains

As organisms with nervous systems have evolved over time to become more well-suited to their environments, the complexity and scale of their nervous systems have evolved along with them. However the relationships between organism size, nervous system complexity, and cognitive ability are challenging to describe and understand. For instance *C. elegans* (the worm), with only 302 neurons in the hermaphrodite (Hobert, 2010), is capable of sensing a range of external stimuli, including temperature, pheromones, attractive/repellent odors, and other chemicals, synthesizing these stimuli and engaging in complex locomotive and mating behaviours in response to them (Barr and Garcia, 2006; Tsalik and Hobert, 2003). *Drosophila melanogaster* (the fruit fly), on the other hand, has nearly 100,000 neurons in its adult brain (Chiang et al., 2011; Scheffer et al., 2020). This more than 300X increase in scale over the worm imparts some of the added stimulus sensitivity and behavioral complexity observed in the fruit fly. Likewise, vertebrates such as *Mus musculus* (mice) and *Homo sapiens* (humans) have developed central brains with 71 million (Keller et al., 2018) and 86 billion (Herculano-Houzel, 2012) neurons respectively, in order to facilitate ever more complex sensory, behavioral, and higher-order cognitive functions.

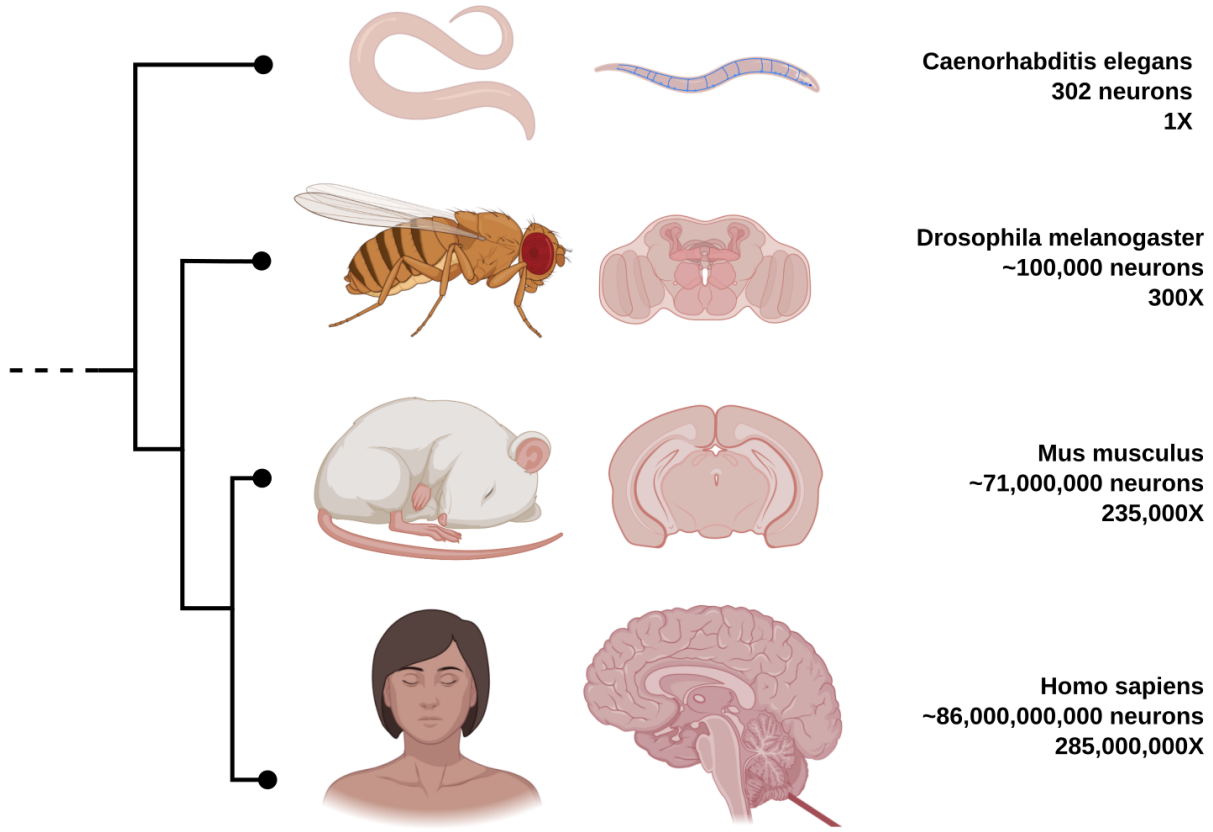

Fig. 1.1: The evolutionary scaling of neurogenesis
Abridged phylogenetic tree (Letunic and Bork, 2021) showing evolutionary lineage relationships between 4 selected model organisms and humans. The number of neurons in adults of each species grows dramatically across evolution, indicated in part by the relative scaling factor (#X) of the number of neurons in each organism vs. *C. elegans*, a relatively simple benchmark organism.

Neurogenesis mechanisms employed across model organisms

A variety of neurogenesis mechanisms have evolved in order to facilitate neurogenesis on these vastly varying scales. Some organisms, such as *C. elegans*, largely employ *direct neurogenesis*, whereby a neural stem cell divides and births a cell that will mature and take on a neural fate without dividing again (Hobert, 2010; White et al., 1986). Though this is perhaps the simplest mechanism possible for generating neurons during development, such a 1-to-1 pairing of neural stem cells to neurons does not scale well to larger nervous systems.

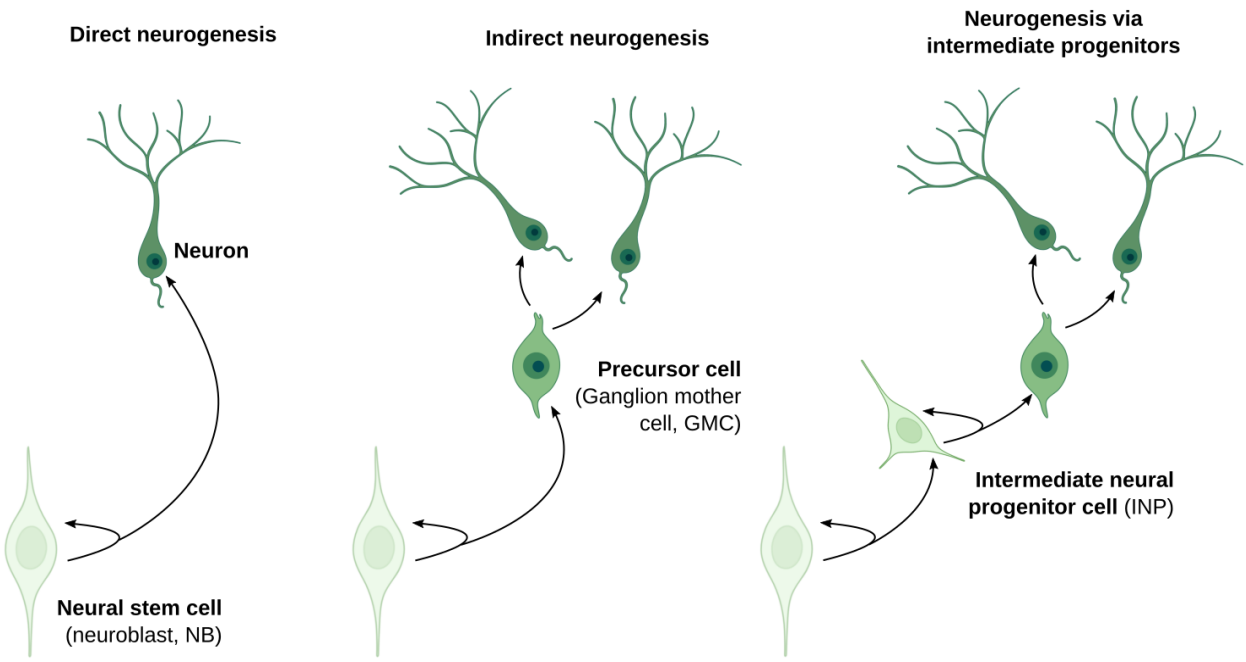

Fig. 1.2: Three common neurogenesis mechanisms
Cell division products are shown at the ends of arrows, with self-directed arrows indicating self-renewal of the mother cell. Common names of important cell types are provided in bold. Names of these cells in the developing *Drosophila* central nervous system are provided in parenthesis. The type I neuroblasts in *Drosophila* predominantly employ indirect neurogenesis and are far more numerous (approximately 100 in each brain lobe) than the type II neuroblasts, which employ neurogenesis via intermediate progenitors and are more scarce (8 in each brain lobe).

More complex organisms, such as *Drosophila melanogaster*, begin to predominantly employ a more complex *indirect neurogenesis* mechanism, whereby a neural stem cell divides and births a *precursor cell* (in *Drosophila*, this is termed a Ganglion Mother Cell or GMC) (Boone and Doe, 2008; Doe, 1992; Matsuzaki et al., 1992). This precursor will mature and divide exactly one more time, giving rise to two daughter neurons/glia, depending primarily on the neural stem cell's age and lineage identity (Homem and Knoblich, 2012). This now 1-to-2 pairing of neural stem cells to neurons may scale better for larger nervous systems, but is still linearly dependent upon the number of neural stem cells present during development and thus poses challenges for organisms whose nervous systems are nearly 285 million times as large as that of *C. elegans*.

In order to break beyond this linear relationship, organisms have developed neural stem cells that generate *progenitors* that can divide multiple times before terminally differentiating. These intermediate progenitor cells (intermediate neural progenitors/INPs in *Drosophila (Boone and Doe, 2008)*, outer radial glia cells/oRGCs in the mouse/human (Hansen et al., 2010; Wang et al., 2011)) maintain their stemness without becoming tumorigenic through a combination of genetic (Bayraktar et al., 2010; Janssens et al., 2014) and metabolic (Bonnay et al., 2020) controls, and can generate many precursor cells throughout their lifetime. Effectively, the use of intermediate neural progenitors enables an exponential increase in the size of any given neural stem cell's neural progeny.

Progenitor patterning as a mechanism for neural fate specification

Regardless of organism size and complexity (at least across the organisms discussed here in this work), neurons generated during development are not identical clones of one another, but rather adopt a variety of unique molecular, morphological, connetomic, and electrophysiological characteristics (Kepecs and Fishell, 2014). Though these characteristics are not necessarily independent of one another, the combination of these and other observable neural characteristics allows us to define a 'neural subtype' or 'neural fate' for each neuron, and subsequently allows us to ask the question "What factors led to this specific neuron acquiring this specific neural fate during development?"

A variety of factors play a role in neural fate specification. At a high level, neural stem cell identity (i.e. which mother cell did this neuron come from?) and developmental time (i.e. how many cell division cycles has this neuron's mother cell gone through prior to this neuron's birth?) have both been observed to play a role in neural fate specification across different organisms (Homem and Knoblich, 2012; Jessell, 2000; Li et al., 2013; Nadadhur et al., 2018; Sulston, 1983). It is generally understood that these two factors impart specific neural fates on their progeny in part via the expression of unique combinations of transcription factors that vary across space and developmental time.

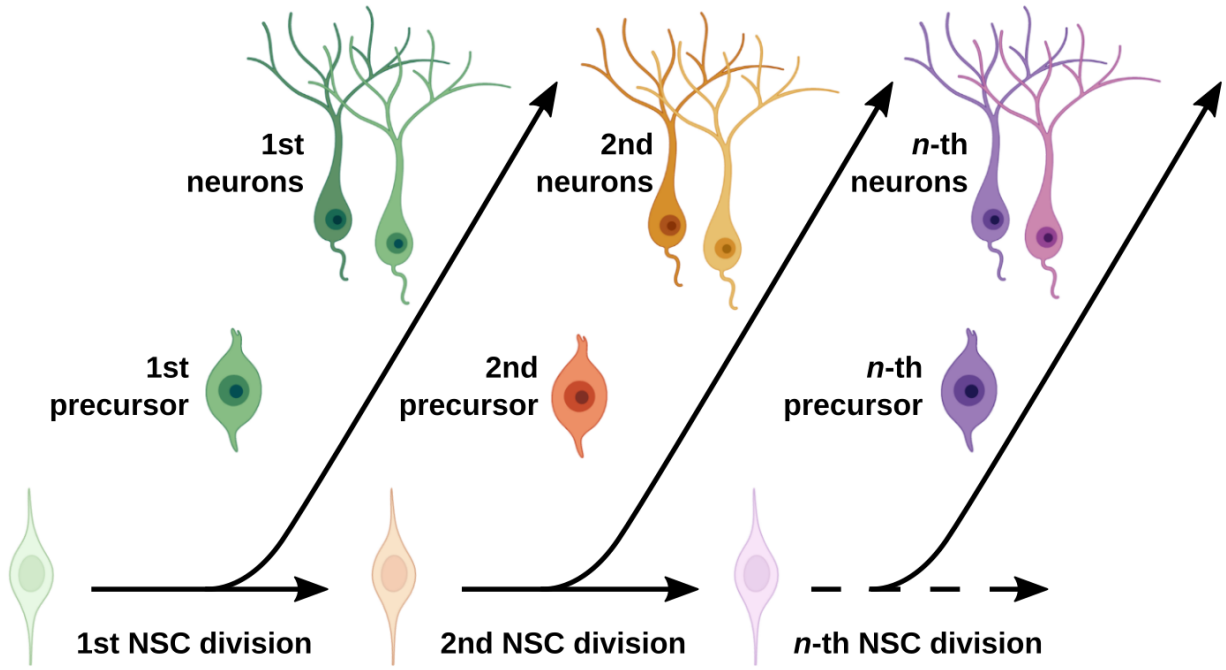

Fig. 1.3: A naive view of neural fate patterning
Simple neurogenesis mechanisms such as direct and indirect (pictured here) neurogenesis yield a 1-to-1 or 1-to-2 pairing of neurons to neural stem cells (NSCs). As these NSCs age and undergo cell division cycles, the temporally varying transcription factors they express change, patterning their neural progeny and ultimately defining their neural fate. Color here indicates the molecularly-defined fate/state of each cell, changing based primarily on the *NSC's* cell division number. Note that absolute developmental time is not to scale in this figure.

When considering neurogenesis via intermediate neural progenitors, we must also consider a third high-level factor - the age of a specific progenitor cell/the number of cell division cycles a specific progenitor cell has undergone before birthing a given neuron. As neural stem cells age and express a variety of neural patterning factors, so to do intermediate progenitors age, expressing their own cascade of temporally varying transcription factors that pattern their neural progeny (Bayraktar and Doe, 2013; Wang et al., 2014) in a combinatorial fashion with neural stem cell age and lineage identity.

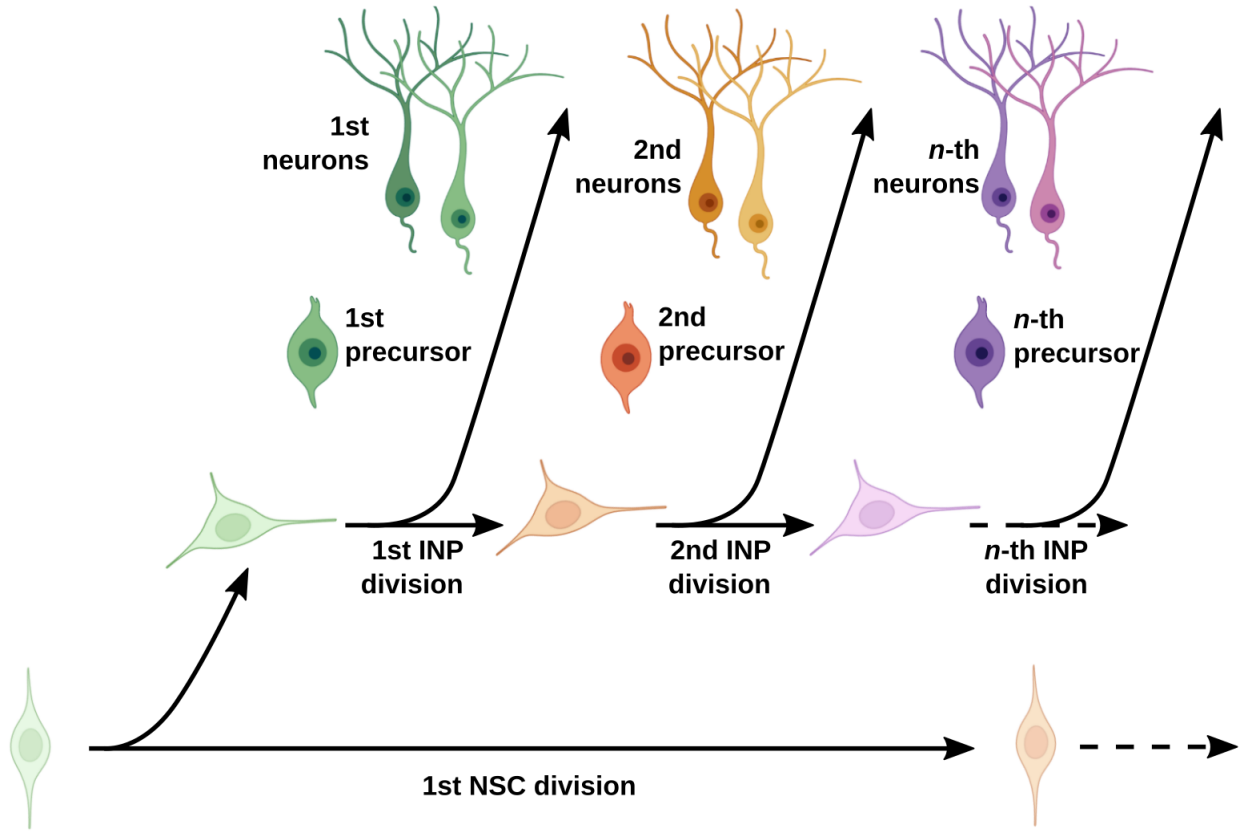

Fig. 1.4: A more complete view of neural fate patterning
Neurogenesis via intermediate neural progenitors enables massive neural diversity by intersecting the fate patterning mechanisms of the neural stem cell (NSC) cell division number with that of their daughter intermediate neural progenitors' (INP) cell division cycles. As NSCs age/divide, they birth progenitors that are patterned with the NSC's temporally varying transcription factors. INPs likewise birth precursor cells that are patterned both by the NSC's and INP's *independent* temporally varying transcription factors, making up a combinatorial neural fate patterning code. Color here indicates the molecularly-defined fate/state of each cell, changing based primarily on the *INP's* cell division number. Note that absolute developmental time is not to scale in this figure.

Additionally, extracellularly expressed spatial patterning cues can induce axonal/dendritic targeting programs that specify a neuron's connectomic fate (Ming et al., 2002; Sanes and Lichtman, 2001). However, this fate specification mechanism generally relies on cells exclusive of a given neuron's lineage (i.e. neurons may be induced to migrate towards/form connections with neurons far away from their mother stem cell lineage). Though not the focus of this work, this mechanism is undoubtedly

part of the broader neural fate specification process, further increasing its complexity.

Drosophila melanogaster as a model for vertebrate neurogenesis

Drosophila melanogaster represents a model organism that recapitulates many features of vertebrate neurogenesis. Unlike the abundant type-I neuroblasts (NB, neural stem cells), the 16 type II NBs in the *Drosophila* brain adopt a neurogenesis process that is directly analogous to that observed in mammalian cortical development (Homem and Knoblich, 2012). During development, each type II NB undergoes repeated asymmetric cell divisions to generate an NB and a sibling progeny that acquires a progenitor identity (i.e. intermediate neural progenitor, INP). Each INP undergoes limited rounds of asymmetric cell division to re-generate and to produce a ganglion mother cell (GMC), which divides once more to become two neuron(s) and/or glial cell(s). Along this NB-INP-GMC-neuron maturation process, cells express a well-defined cascade of transcription factors that mark these cell differentiation stages (Ren et al., 2017; Syed et al., 2017). In parallel, INPs born in each division cycle may express a cascade of transcription factors unique to each NB lineage that contribute to the generation of different neural progenies (Bayraktar and Doe, 2013). It is highly plausible that the combination of these two transcription factor cascades alongside a third molecular axis, which defines unique NBs (i.e., each NB generates a distinct lineage), brings about the generation of a highly diverse neuronal pool.

Characterization of temporally varying molecular factors that pattern intermediate neural progenitors and their progeny

Previous work that revealed the existence of temporally varying patterning mechanisms in the *Drosophila* type II NB lineages relied heavily on *in situ* antibody screening experiments, whereby tissues are stained by a library of antibodies that may or may not be expressed in the cell lineages of interest. Because the type II NB progenies typically do not migrate during the larval stages of development, there exists a spatial relationship between the mother NB, her daughter INPs, and their daughter GMCs/neurons/glia - essentially, as the NB divides, its daughter INP pushes previously

born cells further from the mother NB, and likewise for the INP progenies. In this way, when screening many antibodies at a single point in developmental time, the spatial localization of the antibody signal with respect to the mother NB and INPs gives clues as to the temporal expression dynamics of the protein of interest (Bayraktar and Doe, 2013; Bayraktar et al., 2010). Putative temporally varying genes can be further interrogated for functional importance in *Drosophila* by utilizing the wide range of genetic tools made available by the *Drosophila* research community, most notably the many RNAi lines (Perkins et al., 2015) that can, under proper genetic control, knock-down expression of single genes at the mRNA level in specific sub-populations of cells.

This combination of tools was most famously (in this context) used by Bayraktar and Doe in their 2013 Nature paper, wherein they described the *Dichaete (D), grainy-head (grh), eyeless (ey)* temporally varying transcription factor expression pattern in the type II NBs. They show that as INPs are born they typically express *D* and birth neural progeny expressing *bsh* and/or *D*. As they mature, these INPs stop expressing *D* and begin to express *grh* and eventually *ey*, birthing unique neural progeny at each stage of this INP aging process. They functionally validate the necessity of these genes for generating these neural progeny, and further characterize the *in situ* expression of 52 genes using their antibody library, identifying which (if any) type II NB lineages express these genes at the NB, INP, or neural progeny level.

Studies with similar sets of tools have been performed in the developing murine neural crest by many groups in order to address similar questions of neural fate specification (Simões-Costa and Bronner, 2015) in this much more complex organism. Marker genes that identify key neural fates/progenitor patterning mechanisms have certainly been identified, though the central question still remains: *how many genes/gene transitions are needed to fully specify each neural fate in a single neural stem cell lineage?*

The need for high-throughput molecular screens for more complete characterization of temporally varying neurogenesis mechanisms

Despite the massive significance of previous studies in this space, their reliance on antibody libraries limits their scope. *Drosophila melanogaster* has a genome with more than 14,000 protein-coding genes, at least 700 of which exhibit transcription factor activities (Shokri et al., 2019). Though direct-screening techniques are powerful, the advent of high throughput single-cell mRNA sequencing (scRNA-seq) technologies has enabled researchers to much more broadly investigate the mRNA expression landscape of hundreds of thousands of cells (Macosko et al., 2015; Ziegenhain et al., 2017). Coupled with a vast array of analytical tools that enable us to take these high-dimensional datasets and identify significant molecular signatures from them (Butler et al., 2018; Wolf et al., 2018), researchers can make hypotheses about the number of unique cellular subtypes in the brain (Cocanougher et al., 2019; Saunders et al., 2018), what the functions of these subtypes might be (Ren et al., 2019), and what subtypes might arise together along a common developmental pathway (Cao et al., 2019; Qiu et al., 2017; Soldatov et al., 2019). This combination of experimental and analytical advances has removed the need to rely solely on limited and biased single-molecule screens, instead enabling experiments to yield information about the near-complete mRNA expression landscape of thousands of cells at a time without the need for prior knowledge about/reagents for probing specific genes of interest.

Leveraging genetic tools to improve high-throughput scRNA-seq sensitivity

The most ‘straightforward’ scRNA-seq experimental setup traditionally might involve a researcher collecting a tissue of interest (for example, the whole larval *Drosophila* brain), chemically and mechanically dissociating it into a viable single-cell suspension, and finally processing this entire single-cell suspension using one of many different scRNA-seq mRNA capture techniques; commonly 10X Chromium, DropSeq (Macosko et al., 2015), SeqWell (Gierahn et al., 2017), Smart-seq (Hagemann-Jensen et al., 2020; Picelli et al., 2014), and inDrops (Klein et al., 2015; Lan et al., 2016), though

many other techniques have been and are being developed to accomplish the central goal of capturing and barcoding poly-adenlyated mRNA transcripts from many cells in parallel. I refer to this experimental setup as a “cell atlas” style of scRNA-seq experiment, where a tissue of interest is used in an unbiased manner to characterize (‘build an atlas of’) molecular profiles of cells within that tissue. These studies are broad in their scope and reveal an incredible amount of cell type heterogeneity irrespective of tissue type (see, for example, (Allen et al., 2020; Cao et al., 2019; Cocanougher et al., 2019)).

While “cell atlas” style scRNA-seq datasets effectively characterize the transcriptomes of the majority of cells from a region of interest, cell populations that are classically clustered together (through in situ and/or functional analyses, for example) may not be identified by blind in silico cluster analysis (Kiselev et al., 2019). In addition, broad scRNA-seq studies often do not take advantage of the extensive collection of genetic labelling tools that can highlight classically clustered cell populations, enabling them to be studied in greater detail at far lower cost. For instance, a targeted approach to scRNA-seq is required if we are to confidently and efficiently describe nuanced developmental systems, such as the specification of unique neural subtypes derived from the type-II NB lineages of Drosophila, where inclusion of non-type-II derived cells (making up the majority of the fly brain) would introduce overwhelming noise and confound our analysis, or require an enormous increase in experimental scale and cost.

In order to focus on specific cell populations of interest (for instance, in this work, the *Drosophila* type II NBs and their progeny), transgenic tools such as the UAS-Gal4 (Brand and Perrimon, 1993) system in *Drosophila* and the Cre-lox (Orban et al., 1992) system in mice and other vertebrates can be used to fluorescently label cells that express genes under the control of specific promoter/enhancer sequences. *Drosophila melanogaster* has a famously large library of Gal4 driver lines generated by the research community over the past 3 decades, with more than 7945 Gal4 lines available from the US-based Bloomington Drosophila Stock Center alone (Indiana University,

Bloomington Indiana, USA). Each of these Gal4 driver lines can be crossed to a UAS-reporter/effector line in order to label/effect only those cells expressing Gal4. For example, crossing an Act5C-Gal4 driver line (BDSC stock no. 3954) to a UAS-EGFP reporter line (BDSC stock no. 5428) would label all cells that express *Actin* with *EGFP*. Changing the driver line in the previous example to nSyb-Gal4 (BDSC stock no. 51635) would instead label all neurons (i.e. cells that express the gene *nSyb*) with EGFP.

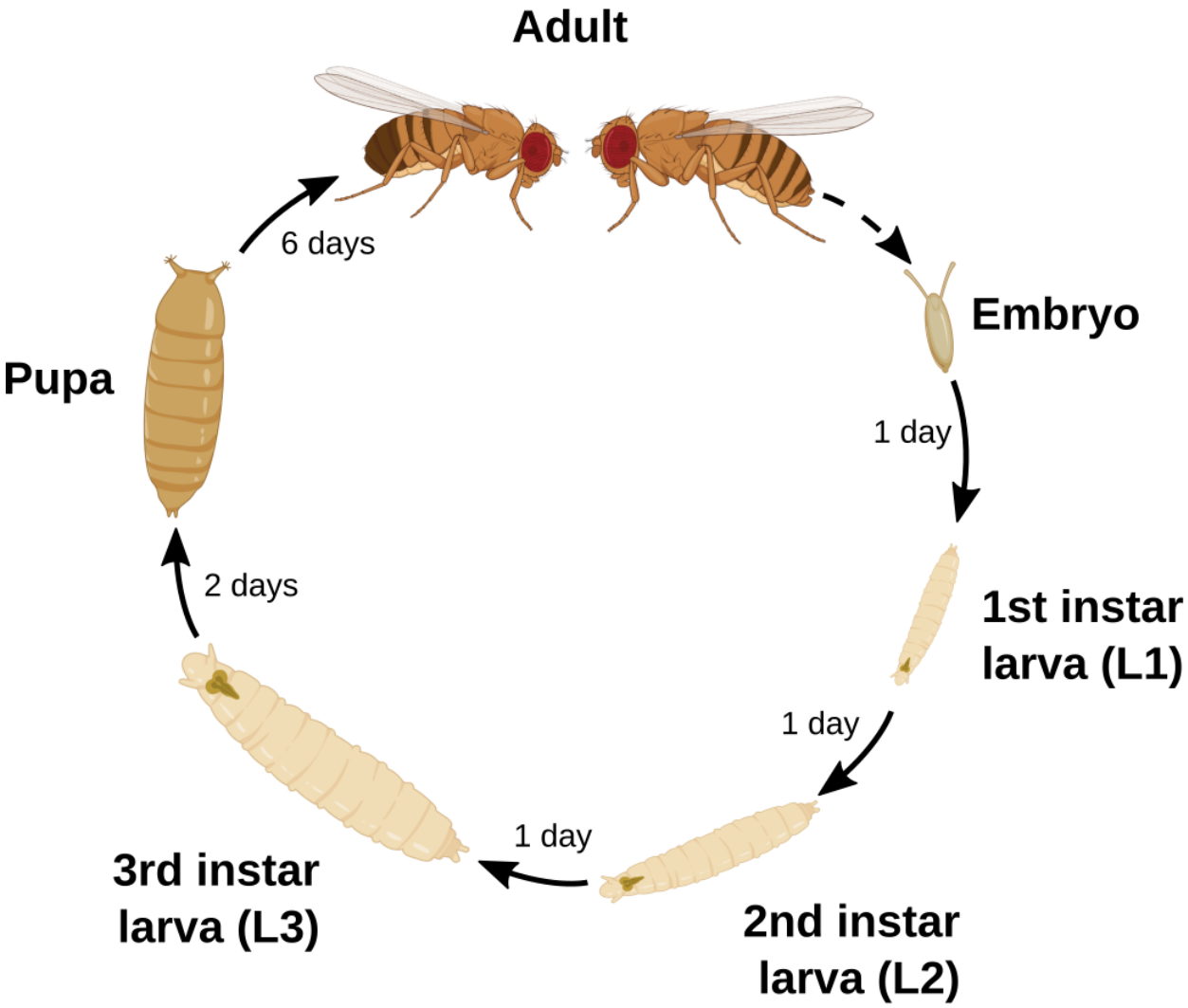

Fig. 1.5: The life cycle of *Drosophila melanogaster*
Arrows with time indicate approximate time for transition from one state to the next, assuming flies are reared at 25C. This rapid developmental time can be shortened/extended by increasing/decreasing the rearing temperature. Total developmental time: 9 days, 11 days, 13 days, 21 days at 29C, 25C, 22C, and 18C, respectively. Physical size not to scale.

In this work, I aimed to perform scRNA-seq on the cells derived from the *Drosophila* type II NB progenies at the late third-instar larval stage of development. Previous work has shown that the gene *earmuff* (*erm*) is required for the generation of type II INPs (Janssens et al., 2014; Koe et al., 2014; Li et al., 2017; Weng et al., 2010), and a fragment of the *erm* promoter known as the (R)9D11 enhancer can be used to label INPs and their progeny when used to drive expression of Gal4 protein (Bayraktar et al., 2010). Using the R9D11-Gal4 driver line (BDSC stock no. 40731) therefore makes it possible to label the nearly 2000 type II derived cells at the late third instar larval stage with a fluorescent reporter, which further enables sorting these cells apart from the far more numerous type I derived cells in larval CNS at this stage (Brunet Avalos et al., 2019; Cocanougher et al., 2019) using fluorescent activated cell sorting (FACS) (Bonner et al., 1972; Herzenberg et al., 2002).

References

Allen, A.M., Neville, M.C., Birtles, S., Croset, V., Treiber, C.D., Waddell, S., and Goodwin, S.F. (2020). A single-cell transcriptomic atlas of the adult Drosophila ventral nerve cord. Elife *9*.

Barr, M.M., and Garcia, L.R. (2006). Male mating behavior. WormBook 1–11.

Bayraktar, O.A., and Doe, C.Q. (2013). Combinatorial temporal patterning in progenitors expands neural diversity. Nature *498*, 449–455.

Bayraktar, O.A., Boone, J.Q., Drummond, M.L., and Doe, C.Q. (2010). Drosophila type II neuroblast lineages keep Prospero levels low to generate large clones that contribute to the adult brain central complex. Neural Dev. *5*, 26.

Bonnay, F., Veloso, A., Steinmann, V., Köcher, T., Abdusselamoglu, M.D., Bajaj, S., Rivelles, E., Landskron, L., Esterbauer, H., Zinzen, R.P., et al. (2020). Oxidative Metabolism Drives Immortalization of Neural Stem Cells during Tumorigenesis. Cell *182*, 1490-1507.e19.

Bonner, W.A., Hulett, H.R., Sweet, R.G., and Herzenberg, L.A. (1972). Fluorescence activated cell sorting. Rev. Sci. Instrum. *43*, 404–409.

Boone, J.Q., and Doe, C.Q. (2008). Identification of Drosophila type II neuroblast lineages containing transit amplifying ganglion mother cells. Dev. Neurobiol. *68*, 1185–1195.

Brand, A.H., and Perrimon, N. (1993). Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. Development *118*, 401–415.

Brunet Avalos, C., Maier, G.L., Bruggmann, R., and Sprecher, S.G. (2019). Single cell transcriptome atlas of the Drosophila larval brain. Elife *8*.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502.

Chiang, A.-S., Lin, C.-Y., Chuang, C.-C., Chang, H.-M., Hsieh, C.-H., Yeh, C.-W., Shih, C.-T., Wu, J.-J., Wang, G.-T., Chen, Y.-C., et al. (2011). Three-dimensional reconstruction of brain-wide wiring networks in Drosophila at single-cell resolution. Curr. Biol. *21*, 1–11.

Cocanougher, B.T., Wittenbach, J.D., Long, X., Kohn, A.B., Norekian, T.P., Yan, J., Colonell, J., Masson, J.-B., Truman, J.W., Cardona, A., et al. (2019). Comparative

single-cell transcriptomics of complete insect nervous systems. BioRxiv.

Doe, C.Q. (1992). Molecular markers for identified neuroblasts and ganglion mother cells in the Drosophila central nervous system. Development *116*, 855–863.

Gierahn, T.M., Wadsworth, M.H., Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C., and Shalek, A.K. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat. Methods *14*, 395–398.

Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A.J.M., Faridani, O.R., and Sandberg, R. (2020). Single-cell RNA counting at allele and isoform resolution using Smart-seq3. Nat. Biotechnol. *38*, 708–714.

Hansen, D.V., Lui, J.H., Parker, P.R.L., and Kriegstein, A.R. (2010). Neurogenic radial glia in the outer subventricular zone of human neocortex. Nature *464*, 554–561.

Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. Proc Natl Acad Sci USA *109 Suppl 1*, 10661–10668.

Herzenberg, L.A., Parks, D., Sahaf, B., Perez, O., Roederer, M., and Herzenberg, L.A. (2002). The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. Clin. Chem. *48*, 1819–1827.

Hobert, O. (2010). Neurogenesis in the nematode Caenorhabditis elegans. WormBook.

Homem, C.C.F., and Knoblich, J.A. (2012). Drosophila neuroblasts: a model for stem cell biology. Development *139*, 4297–4310.

Janssens, D.H., Komori, H., Grbac, D., Chen, K., Koe, C.T., Wang, H., and Lee, C.-Y. (2014). Earmuff restricts progenitor cell potential by attenuating the competence to respond to self-renewal factors. Development *141*, 1036–1046.

Jessell, T.M. (2000). Neuronal specification in the spinal cord: inductive signals and transcriptional codes. Nat. Rev. Genet. *1*, 20–29.

Keller, D., Erö, C., and Markram, H. (2018). Cell densities in the mouse brain: A systematic review. Front. Neuroanat. *12*, 83.

Kepecs, A., and Fishell, G. (2014). Interneuron cell types are fit to function. Nature *505*, 318–326.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Koe, C.T., Li, S., Rossi, F., Wong, J.J.L., Wang, Y., Zhang, Z., Chen, K., Aw, S.S., Richardson, H.E., Robson, P., et al. (2014). The Brm-HDAC3-Erm repressor complex suppresses dedifferentiation in Drosophila type II neuroblast lineages. Elife *3*, e01906.

Lan, F., Haliburton, J.R., Yuan, A., and Abate, A.R. (2016). Droplet barcoding for massively parallel single-molecule deep sequencing. Nat. Commun. *7*, 11784.

Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res.

Li, X., Erclik, T., Bertet, C., Chen, Z., Voutev, R., Venkatesh, S., Morante, J., Celik, A., and Desplan, C. (2013). Temporal patterning of Drosophila medulla neuroblasts controls neural fates. Nature *498*, 456–462.

Li, X., Chen, R., and Zhu, S. (2017). bHLH-O proteins balance the self-renewal and differentiation of Drosophila neural stem cells by regulating Earmuff expression. Dev. Biol. *431*, 239–251.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Matsuzaki, F., Koizumi, K., Hama, C., Yoshioka, T., and Nabeshima, Y. (1992). Cloning of the Drosophila prospero gene and its expression in ganglion mother cells. Biochem. Biophys. Res. Commun. *182*, 1326–1332.

Ming, G., Wong, S.T., Henley, J., Yuan, X., Song, H., Spitzer, N.C., and Poo, M. (2002). Adaptation in the chemotactic guidance of nerve growth cones. Nature *417*, 411–418.

Nadadhur, A.G., Leferink, P.S., Holmes, D., Hinz, L., Cornelissen-Steijger, P., Gasparotto, L., and Heine, V.M. (2018). Patterning factors during neural progenitor induction determine regional identity and differentiation potential in vitro. Stem Cell Res. *32*, 25–34.

Orban, P.C., Chui, D., and Marth, J.D. (1992). Tissue- and site-specific DNA recombination in transgenic mice. Proc Natl Acad Sci USA *89*, 6861–6865.

Perkins, L.A., Holderbaum, L., Tao, R., Hu, Y., Sopko, R., McCall, K., Yang-Zhou, D., Flockhart, I., Binari, R., Shim, H.-S., et al. (2015). The transgenic rnai project at harvard medical school: resources and validation. Genetics *201*, 843–852.

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Ren, Q., Yang, C.-P., Liu, Z., Sugino, K., Mok, K., He, Y., Ito, M., Nern, A., Otsuna, H., and Lee, T. (2017). Stem Cell-Intrinsic, Seven-up-Triggered Temporal Factor Gradients Diversify Intermediate Neural Progenitors. Curr. Biol. *27*, 1303–1313.

Sanes, J.R., and Lichtman, J.W. (2001). Induction, assembly, maturation and maintenance of a postsynaptic apparatus. Nat. Rev. Neurosci. *2*, 791–805.

Scheffer, L.K., Xu, C.S., Januszewski, M., Lu, Z., Takemura, S.-Y., Hayworth, K.J., Huang, G.B., Shinomiya, K., Maitlin-Shepard, J., Berg, S., et al. (2020). A connectome and analysis of the adult Drosophila central brain. Elife *9*.

Shokri, L., Inukai, S., Hafner, A., Weinand, K., Hens, K., Vedenko, A., Gisselbrecht, S.S., Dainese, R., Bischof, J., Furger, E., et al. (2019). A Comprehensive Drosophila melanogaster Transcription Factor Interactome. Cell Rep. *27*, 955-970.e7.

Simões-Costa, M., and Bronner, M.E. (2015). Establishing neural crest identity: a gene regulatory recipe. Development *142*, 242–257.

Sulston, J.E. (1983). Neuronal cell lineages in the nematode Caenorhabditis elegans. Cold Spring Harb. Symp. Quant. Biol. *48 Pt 2*, 443–452.

Syed, M.H., Mark, B., and Doe, C.Q. (2017). Steroid hormone induction of temporal gene expression in Drosophila brain neuroblasts generates neuronal and glial diversity. Elife *6*.

Tsalik, E.L., and Hobert, O. (2003). Functional mapping of neurons that control locomotory behavior in Caenorhabditis elegans. J. Neurobiol. *56*, 178–197.

Wang, X., Tsai, J.-W., LaMonica, B., and Kriegstein, A.R. (2011). A new subtype of progenitor cell in the mouse embryonic neocortex. Nat. Neurosci. *14*, 555–561.

Wang, Y.-C., Yang, J.S., Johnston, R., Ren, Q., Lee, Y.-J., Luan, H., Brody, T., Odenwald, W.F., and Lee, T. (2014). Drosophila intermediate neural progenitors produce lineage-dependent related series of diverse neurons. Development *141*, 253–258.

Weng, M., Golden, K.L., and Lee, C.-Y. (2010). dFezf/Earmuff maintains the restricted developmental potential of intermediate neural progenitors in Drosophila. Dev. Cell *18*, 126–135.

White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. Philos. Trans. R. Soc. Lond. B. Biol. Sci *314*, 1–340.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol. Cell *65*, 631-643.e4.

Chapter 2

Characterizing the Developmental Landscape of the Type-II Neuroblast Lineages of *Drosophila melanogaster*

Overview[1]

In the type-II NB lineages of Drosophila, we set out to broadly classify the molecular factors that define the neural progenies of dividing INPs along three key fate-patterning axes, i.e., differentiation state, division number, and progenitor lineage (Fig. 2.5A) using targeted scRNA-seq. We created a long-living fluorescent reporter to brightly label the type-II progenies at the 3rd instar larval stage and FACS sorted them in preparation for 10X Chromium scRNA-seq (Fig. 2.1). We subsequently recovered transcriptomes containing 11622 genes from 6092 cells. Through an iterative process of cell clustering, marker gene analysis, pseudotime analysis, and in situ validation, we identified genes that vary in expression along all three neural fate-patterning axes mentioned above. These genes include markers that globally define the INP, GMC, and neuron differentiation stages in most NB lineages. Further in silico analysis suggested molecular factors that are uniquely expressed in subpopulations of INPs, GMCs, immature and mature neurons. Subsequent in situ mRNA staining recovered the spatial relationship of these molecular factors, which clarified the cell division number and NB lineage specificity. We finally identified novel markers that exclusively label distinct neural subsets. These new markers further enabled building novel neural developmental trajectories that lead to unique neuronal cell fates. Our multi-informatic

1This chapter largely encompasses our 2021 publication in Cell Report on this topic. Please cite: Michki, N.S., Li, Y., Sanjasaz, K., Zhao, Y., Shen, F.Y., Walker, L.A., Cao, W., Lee, C.-Y., and Cai, D. (2021). The molecular landscape of neural differentiation in the developing Drosophila brain revealed by targeted scRNA-seq and multi-informatic analysis. Cell Rep. 35, 109039.

approach to targeted scRNA-seq experimental design and analysis provides a roadmap for navigating the differentiation process of complex brains. Our annotated scRNA-seq data and interactive analysis tools provide valuable resources for future discoveries.

Dissociation and FACS selection of type-II derived cells

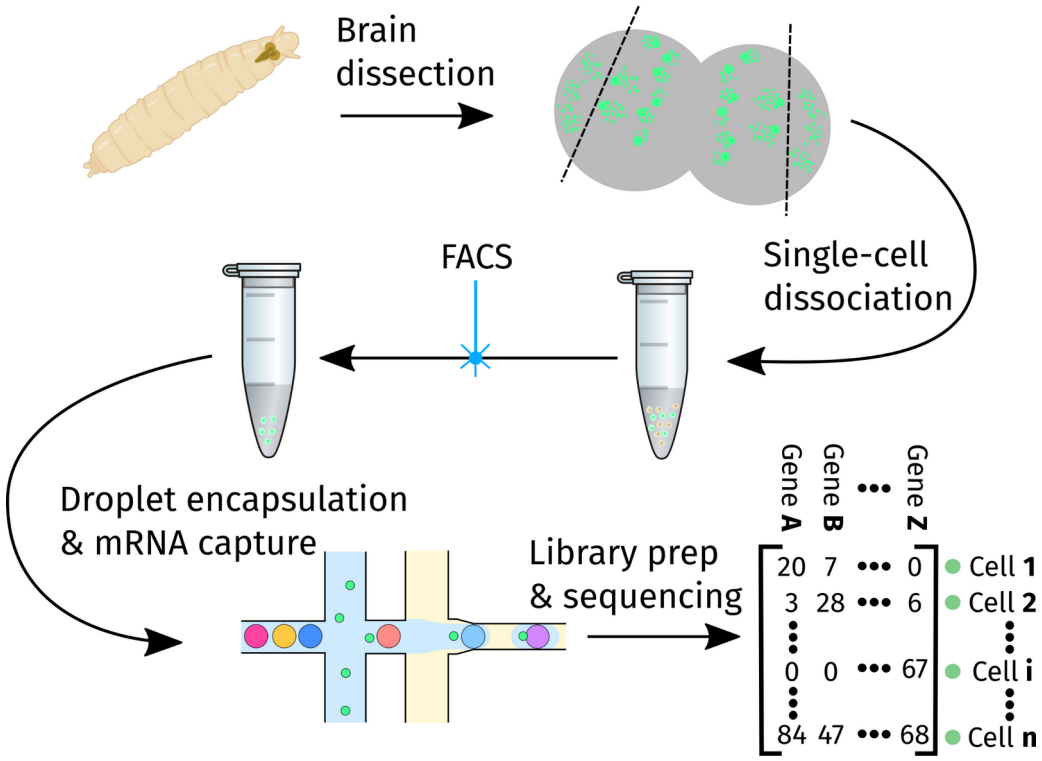

Fig. 2.1: Experimental overview
Late third instar larvae were collected, their brains dissected and dissociated, and the type II cells from them sorted on a FACS machine. These cells were loaded onto a 10X Chromium mRNA capture chip and used to generate a sequencing library for downstream analysis.

[;;R9D11-Gal4/UAS-hH2B::2xmNG] larvae (n=20) were rinsed and their brains dissected using dissection scissors and forceps at the late L3 stage (wandering larvae) in ice cold Rinaldini's solution. These brains were subsequently transferred to a poly-L-lysine coated coverslip that was immersed in Rinaldini's solution, attaching only the VNC to coverslip and leaving the central brain lobes unattached. These brain lobes were then further dissected using a tungsten needle by inserting the needle into each

brain lobe at approximately the midpoint of the lobe and moving the needle laterally. This process removed a lot but not all the cells on the lateral portions of each brain lobe, which includes the developing optic lobe. The remaining OL cells were later excluded from our final scRNA-seq dataset using known marker genes (detailed above).

Dissected brains were transferred to a DNA low-binding 1.5mL tube in 30µL of dissection liquid (Rinaldini's solution) using a p200 pipette equipped with a siliconized p200 tip that was cut and flame-smoothed approximately 1/4 of the way up the tip. The siliconized tips are lower-binding and make it less likely for brains to stick to them. Cutting the tip and smoothing the opening makes it easier for the brains to move into the tip. The 1.5mL tube was pre-filled with 50µL of fresh, cold Rinaldini's solution, and upon transfer of the brains, 10µL of 20mg/mL papain, 10µL of 20mg/mL type-I collagenase, and 1µL of 15µM ZnCl were added to the tube, bringing the total reaction volume to 100µL. The tube was closed and mixed gently by flicking, then incubated on a heat block at 37°C for 1hr. During this incubation, the tube was flicked for mixing at 10min intervals, flicking the tube until the brains were visibly disturbed into the solution. After the 1hr incubation, 2µL of 100µM E-64 solution was added to the mixture to stop the papain digestion. To break down the apparent intact brains, the mixture was triturated at a ~1 Hz frequency for 30 times using a p100 pipette set to 70µL and equipped with an uncut p200 siliconized tip. After the first 5 triturations, the brains should be seen largely dissociated to the naked eye. Further triturations break down the brain completely into single-cell suspensions including the VNC, which is quite resilient to dissociation.

After trituration, the cell suspension was diluted with 400µL Schneider's media + 10% FBS which further quenches the enzymatic digestion and stabilizes the cells. 1µL of DRAQ5 DNA stain (Thermo Fisher Scientific Inc.) was added to label cells apart from debris generated in the dissociation process.

The sorting-ready cell suspension was transferred to a 5mL plastic FACS snap-cap tube

on ice. Cells from non-Gal4 driver brains were dissociated in a similar manner and were sorted first on a Sony MA900 FACS machine to set the gate for using DRAQ5 to separate DNA containing cells from debris and set the gate for non-mNG expressing cells.

Sorted cells were captured in a DNA low-binding 1.5mL tube pre-filled with 100µL of Schneider's media + 10% FBS. Cells were spun down at 400x g for 4 minutes and the solution volume was reduced to 40µL before resuspending by gentle pipetting with a p200 siliconized pipette tip. 5µL of this suspension was removed to count cells using an epifluorescence microscope by plating them in a single well of a 96 well plate, pre-filled with 45µL of Schneider's media + 10% FBS. The rest of the cells were transported on ice to the University of Michigan Advanced Genomics Core and approximately 10,000 cells were loaded for 10X Chromium V3 sequencing following the manufacturer's instruction.

scRNA-sequencing

Two replicate experiments were performed: one on a 10X Chromium v2 chip, and one on a 10X Chromium v3 chip. Input cell counts were approximately equal across replicates.

Approximately 10,000 type-II derived cells were used as input to a single channel of a 10X Chromium chip. The mRNA was subsequently reverse transcribed, amplified, and prepared for sequencing on an Illumina NovaSeq-6000 chip (University of Michigan Advanced Genomics Core). The library was sequenced for a total of 385M paired-end reads with 28bp for the cell barcode and UMI and 110bp for cDNA inserts.

scRNA-seq mapping and downstream analysis

Reads were mapped using both Cell Ranger (for initial analysis) and STAR-solo (for our final analysis, with mNG added to the genome) (Dobin et al., 2013) to the Drosophila

genome assembly provided by ENSEMBL, build BDGP6 (2014-07).

The downstream scRNA-seq analysis was performed using scanpy (Wolf et al., 2018), and our analysis was formalized into the MiCV web tool generated in this work (https://micv.works). In brief, cells were filtered by requiring between 200-4100 unique genes/cell (to exclude debris and some doublets) and genes were filtered by requiring at least 2 cells to express it at greater than 1 UMI/cell. UMI counts were normalized to a total sum of 1e6 counts/cell (conversion to counts-per-million/CPM) and subsequently log-transformed by calculating ln(1+CPM) for each gene for each cell. The top 2000 highly variable genes were identified using the cell-ranger method (Zheng et al., 2017) and these genes were used to perform a principal component analysis (PCA, n=50pcs). As two replicate experiments (batches) needed to be integrated across different sequencing chemistries (10X v2 and v3), the harmony (Korsunsky et al., 2019) data integration algorithm was used to batch-correct this PCA representation of the data before proceeding to neighborhood identification (k=20), and finally a UMAP projection (2D). Clusters were identified using the Leiden algorithm (Traag et al., 2019), an optimized version of the Louvain algorithm (Blondel et al., 2008), with varying clustering resolutions. For most of the type-II only UMAP projections displayed in this work, the clustering resolution was 0.6, with 1 being a standard default (and higher numbers leading to more granular clustering of cells). Marker genes were identified using logistic regression analysis, implemented in scanpy.

Type-II neuroblast derived cells are uniquely identified from the mixed optic lobe cell population using descriptive quality control metrics and clustering

To perform targeted scRNA-seq, we brightly labeled the type-II NB progenies with a long-lasting fluorescent reporter. We created an UAS-hH2B::2xmNG reporter fly, in which two copies of the mNeonGreen (2xmNG) fluorescent protein are fused to the C'-terminus of the human histone 2B protein (hH2B). This leverages the expression of multiple copies of a bright fluorescent protein alongside the slower turn-over rate of the histone protein (Tumbar et al., 2004). To validate labeling fidelity, we crossed UAS-

hH2B::2xmNG to an R9D11-Gal4 driver (Weng et al., 2010). We found that mNG labeled type-II NB progenies and a small subset of medial optic lobe (OL) cells in larval brains (Fig. 2.2A). When comparing our UAS-hH2B::2xmNG to the previously used UAS-IVS-myr::tdTomato reporter, we found that the membrane-targeted myr::tdTomato cells formed clusters that are smaller than the hH2B::2xmNG labeled cells (Fig. 2.2B-D). This indicates that the slow hH2B::2xmNG turnover preserved labeling in progeny cells in which Gal4 was no longer expressed. Finally, the bright nuclear mNG labeling enabled reliable FACS selection for targeted 10x Chromium scRNA-seq (Fig. 2.1).

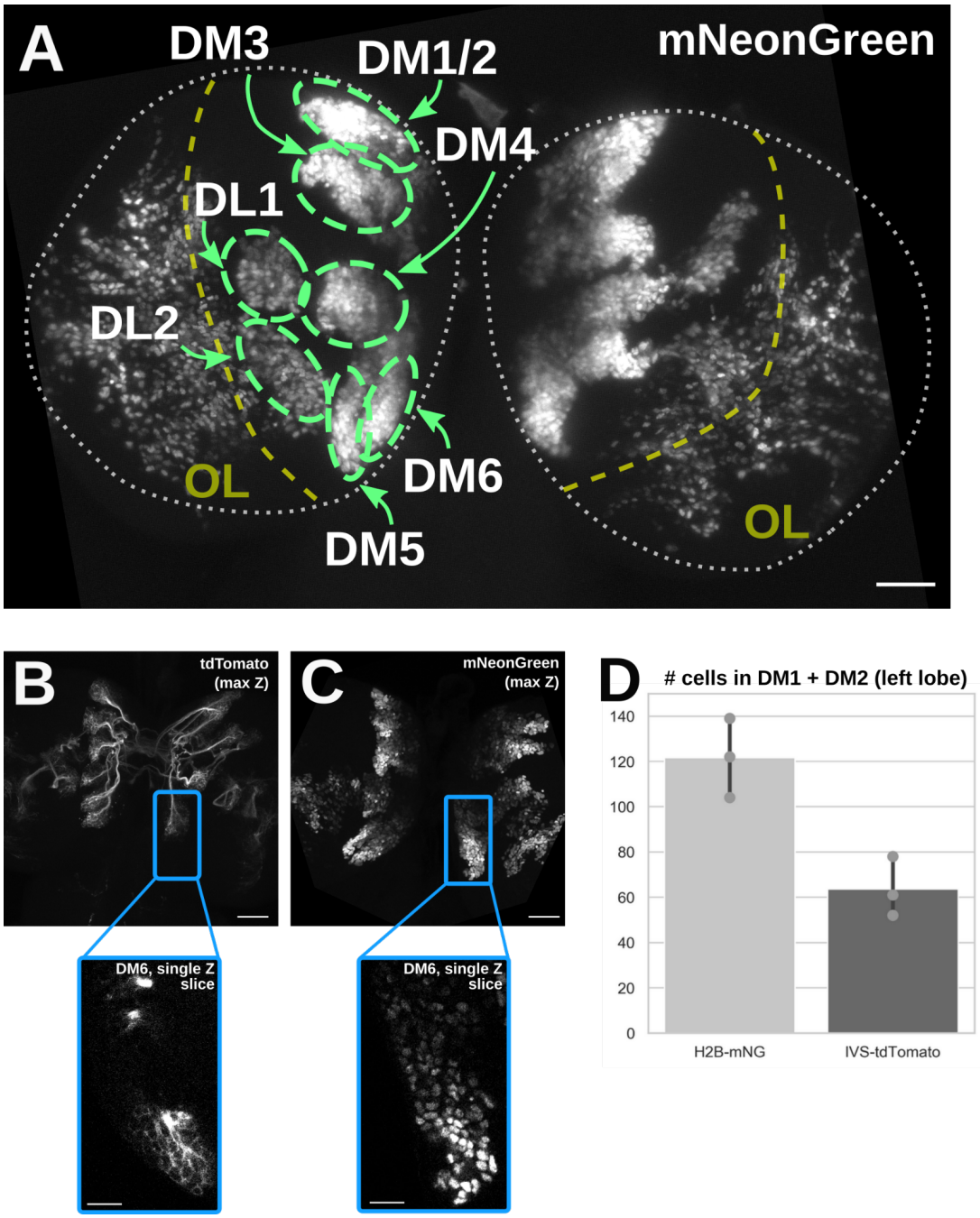

Figure 2.2: A long-lasting nucleus UAS-hH2B::2xFP (mNeonGreen/tagBFP) reporter labels more cells in the type-II progenies than the membrane UAS-IVS-myc::tdTomato reporter

(A) A max-Z projection of the novel UAS-hH2B::2xmNeonGreen reporter driven under the control of R9D11-Gal4 at the third instar larval developmental stage. The type-II progenies are outlined in green dashed lines, and the approximate boundary between the central brain and the developing optic lobe (OL) is marked by the yellow dashed

line. Gamma correction (γ = 0.5) was applied to better visualize the dimmer OL cells. (B) The membrane-bound tdTomato reporter is driven under the control of R9D11-Gal4 and its lineage labeling is compared to that of our (C) nucleus-targeted 2xmNG reporter. (D) Quantifications of labeled cells in clusters DM1 and DM2 in late third instar larvae brains. Bars represent the mean of manual cell counts from DM1 and DM2 in three brains for each genotype; points represent cell counts for the individual replicates. Scale bars, 30 µm in overviews of (A,B,C), 10 µm in insets of (B,C).

Subsequently, we projected the scRNA-seq data onto a 2D UMAP plot and overlaid the counts of all genes, unique transcripts (UMI), and mitochondrial genes as part of routine scRNA-seq quality control (Fig. 2.3).

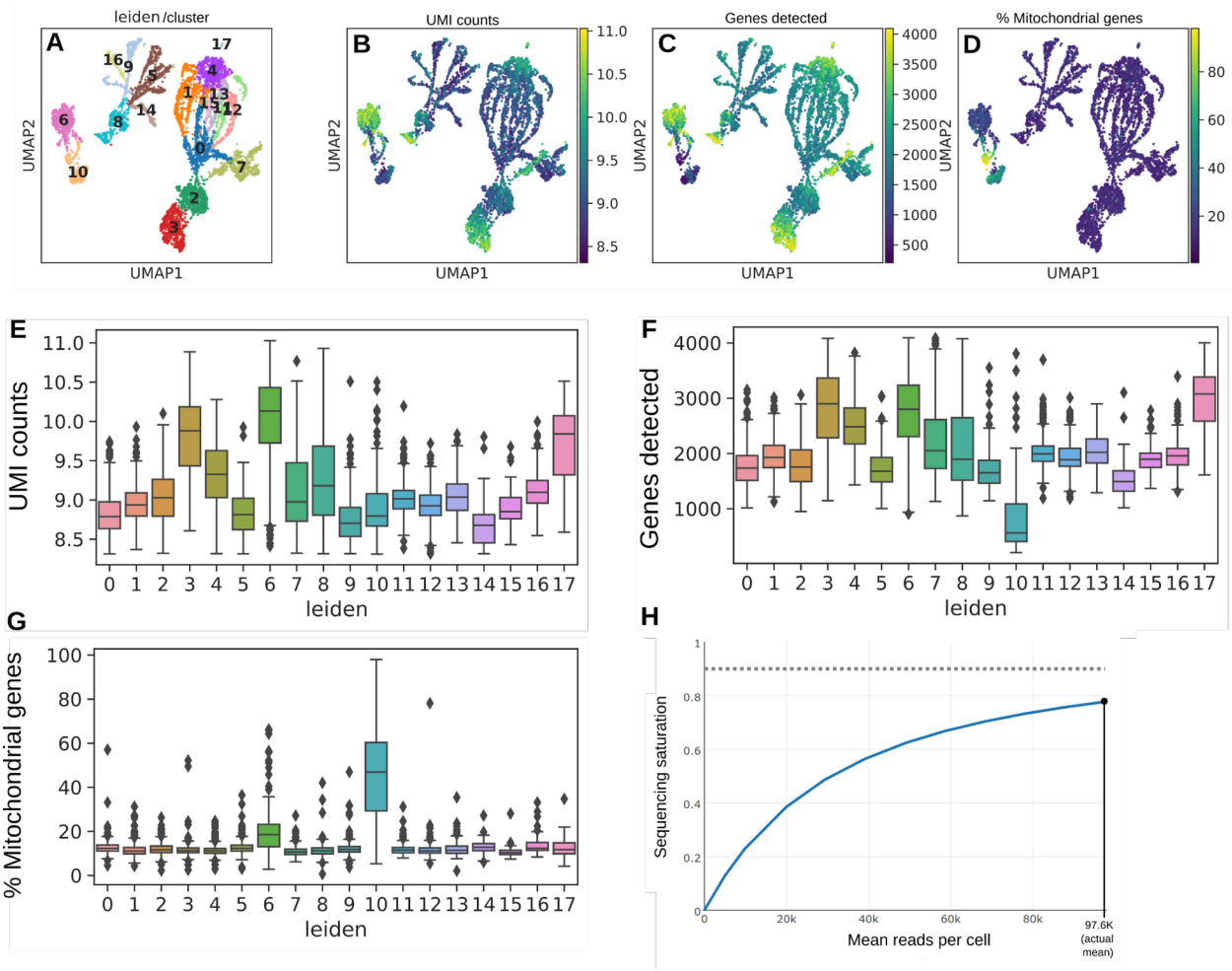

Fig. 2.3: Sequencing QC metrics indicate that captured cells are healthy and diverse in transcriptional activity

(A) UMAP with colorimetric and numerical labels for each automatically assigned cluster (Leiden algorithm, resolution=0.5). (B-D) UMAP of cells overlaid with their ln(1+UMI

counts), number of unique genes detected, and percentage of mitochondrial genes, respectively. (E-G) The QC metrics from above but summarized as boxplots on a per-cluster basis. Clusters 3, 4, and 6 are large groups of cells that have particularly high UMI counts and gene detection rates, indicating that they are transcriptionally very active. Cluster 1 is the group of type-II derived INPs described in this work. Cluster 6 cells are likely glia based on the expression of repo (Fig. 2.6) and cluster 8 cells are likely progenitors in the OL cells based on the expression of CycE (Fig. 2.4). Cluster 4 is a group of maturing neurons that has a higher than average gene detection rate, and strongly expresses Imp (data not shown), an IGF-II RNA-binding protein that is responsible for a number of RNA trafficking functions, notably being required for axonal growth and remodeling (Medioni et al. 2014). As the type-II neuronal progenies extend large axonal bundles across the commissure during development, it is possible that this transcriptionally active group of Imp+ neurons are the ones actively undergoing this process. Boxes represent interquartile range (IQR) of data; midline represents median; whiskers represent range, with the exception of outliers which are represented by points (values are 1.5 times the IQR beyond the low or high quartile). (H) Predicted sequencing saturation curve generated using CellRanger, indicating that at our sequencing depth we have recovered nearly 80% of unique genes that might be found in each cell.

When overlaying the hH2B::2xmNG reporter transcript counts, we found that mNG transcripts were expressed non-uniformly, with pockets of cells expressing the hH2B::2xmNG transcript at a significantly higher level than others in the dataset (Fig. 2.4A). To examine whether this non-uniform expression pattern reflects true biological variance, we performed in situ RNA staining for mNG using the HCRv3 protocol (Choi et al., 2018) and imaged the native mNG fluorescence to compare the relationship of mNG transcripts and proteins (Methods). We found that each of the type-II clusters indeed expresses a high level of mNG transcripts in only a small subpopulation of cells near the tip of each lineage (Fig. 2.4B-D). This spatial localization, coupled with co-expression of mNG transcripts with D in CycE+ cells (data not shown) leads us to conclude that the R9D11 enhancer fragment's expression is tightly restricted to newly-born INPs and their daughter GMCs, emphasizing the need for long-living reporters for investigation of neural subtypes derived from the type-II NBs.

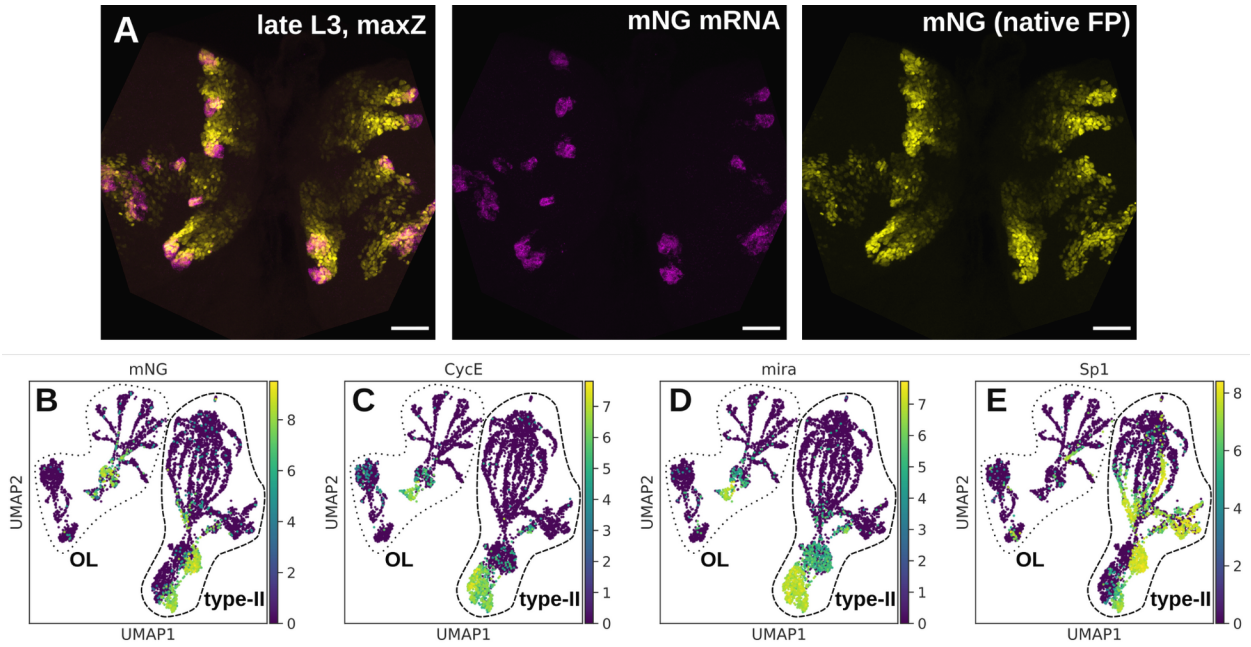

Fig. 2.4: R9D11-Gal4 driven reporter mRNA expression is restricted to a small portion of each type-II lineage

(A) A composite maximum z-projection of a late third instar larval brain expressing our novel UAS-hH2B::2xmNeonGreen (mNG) reporter under the control of R9D11-Gal4, with native mNG fluorescence labeling the type-II progenies and mNG mRNA labeled using our mNG HCR v3 probes. At the tip of each type-II lineage, there is a burst of expression of mNG transcripts (middle panel) that does not persist throughout the lineage but rather remains restricted to what is presumably the youngest mINPs. (B) This assessment is further validated using our scRNA-seq data, wherein we find that mapped mNG transcript expression is multiple log2-fold higher in young mINPs and their daughter GMCs, based on the expression of CycE/mira for mINPs and Sp1 for young mINPs and their progeny (C-E) (see also Fig. 2.9 in the main text). In the optic lobe (the large connected group of cells on the right-hand side of the UMAP projection), the expression is not restricted to cells with the highest CycE/mira expression and so it is possible that the R9D11 enhancer element is active in a non-progenitor population in the optic lobe. Scale bars: 30 μm.

To further ensure the specificity of our analysis to type-II cells, we performed an in silico filtering to exclude the optic-lobe cells that are also labeled by R9D11-Gal4 (Bayraktar et al., 2010). Based on prior literature, at least two genes are not expressed in the developing optic-lobe (lncRNA:cherub and dati; see in situ expression patterns from (Landskron et al., 2018; Schinaman et al., 2014), respectively). In the unsupervised

clustering and UMAP projection, two groups of cells can be clearly separated as cherub+/dati+ and cherub-/dati-, which we define as putative type-II and OL cells, respectively (Fig. 2.5C-E).

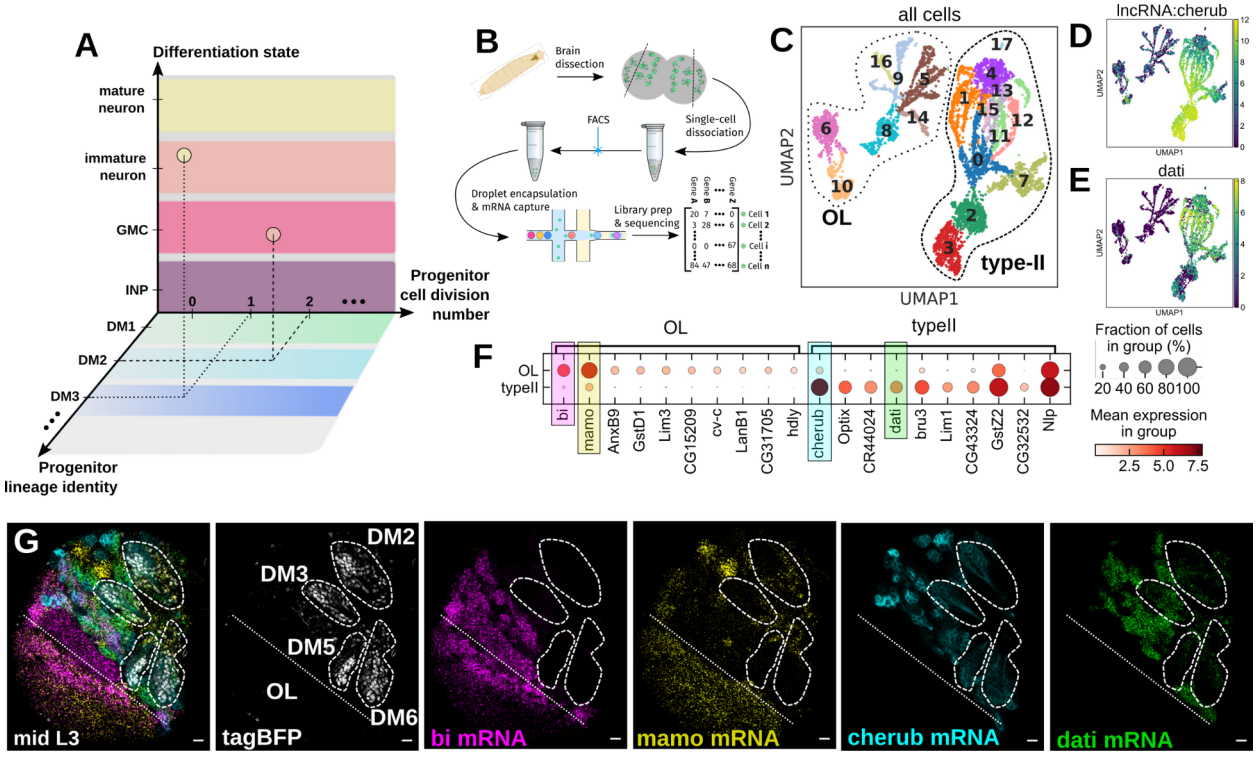

Fig. 2.5: Drosophila type-II neuronal fate specification model, experiment overview, and in silico dissection of the optic lobe and type-II derived cells
(A) A diagram of the major axes that determine cell "state" in this work. Each sequenced cell is defined in part by factors that are specific to the lineage identity, intermediate progenitor cell division number, and differentiation state. (B) Overview of our targeted scRNA-seq experimental strategy. (C) Cells plotted in the first 2 dimensions of a UMAP projection. Color represents automatic cluster assignment by the Leiden algorithm (resolution = 0.5). (D-E) Expression of the long non-coding RNA cherub and the transcription factor dati are known to be exclusive of the optic lobe in 3rd instar larvae. Groups of cells that lack expression of these genes are likely optic lobe cells that also express Gal4 under the control of the R9D11 fragment of the erm promoter. (F) Separating the putative type-II/optic lobe cells into two groups and performing logistic regression analysis reveals genes that are up-regulated between the two. (G) A single z-slice of one brain lobe from the developing (mid L3 stage) larval brain. UAS-hH2B::2xtagBFP is driven under the control of R9D11-Gal4 and marks the type-II lineages, only four of which are visible in this z-slice. lncRNA:cherub and dati mRNA are largely expressed by type-II cells, while bi and mamo mRNA are largely expressed in

the developing optic lobe (boundary marked by the diagonal line). Scale bars: 10um in all images.

To identify other potential marker genes to separate OL and type-II cells, we performed a logistic regression-based marker gene analysis (Ntranos et al., 2018) comparing these two major groups against one another (Fig. 2.5F). The transcription factors mamo and bi are upregulated in the putative OL cells when compared to the putative type-II cells, among others. To confirm this, we generated HCR probes against mamo and bi as novel markers for the OL, and lncRNA:cherub and dati as markers for cells not in the OL. We subsequently stained larval R9D11-hH2B::2xtagBFP brains (Fig. 2.5G), and clearly show that bi and mamo are both predominantly expressed in the OL, and lncRNA:cherub and dati are predominantly excluded from the OL. Why mamo is upregulated in cells in the OL is unknown. Prior work has shown that mamo is required for specification of α'/β' mushroom body neurons in the developing CNS (Liu et al., 2019). Further study of its role in the OL may elucidate novel function there. On the other hand, bi has been shown to be indispensable for neural differentiation in the OL (Pflugfelder et al., 1990). Our finding of bi being excluded from the type-II lineages expands our knowledge of its expression specificity.

From our in silico filtering process, we confidently separated the type-II derived cells from optic lobe cells that were also captured in our scRNA-seq experiment. Only these type-II derived cells were carried forward for our downstream analysis.

Pseudotime analysis describes the continuous differentiation stages of type-II derived cells

Knowing that the R9D11-hH2B::2xmNG reporter specifically labels type-II progenies from INPs to maturing neurons, we aimed to first align each cell along a continuous cellular differentiation state axis (Fig. 2.5A). We expected this would reveal the most prominent underlying structure of our data because, in the case of type-II neurogenesis, all cells will similarly transition through the INP, to GMC, to immature, to mature neuron differentiation states. Using the Markov chain-based pseudotime analysis algorithm

Palantir was a natural choice as Markov chains describe discrete transitions that occur randomly based upon a continuous probability distribution (Setty et al., 2019). Given a properly chosen starting cell, Palantir aligns cells in our scRNA-seq data based upon the path of fewest transcriptomic changes propagating from the starting cell.

Cells expressing high levels of the INP markers CycE and D are good candidate starting cells for Palantir (Bayraktar and Doe, 2013; Yang et al., 2017). To easily identify these cells from the UMAP plot, we built a Multi-informatic Cellular Visualization web tool (MiCV) to display the single cell co-expression pattern of multiple genes in the 2D/3D UMAP plots. Furthermore, users can conveniently select a subset of cells for specific analysis, such as picking the starting cell(s) for Palantir, by combining mouse-click selections from the parallel plots generated by MiCV (Methods). We overlaid the pseudotime result onto the reprojected 2D UMAP plot that only included type-II NB derived cells. Based on the expression of known marker genes (Fig. 2.6), we predicted INP, GMC, immature, and mature neuron clusters (Fig. 2.7A, dash lines). Interestingly, these cell maturation state clusters aligned well with the pseudotime arrangement. For example, using MiCV, we displayed the single cell co-expression pattern of CycE, dap, and nSyb (Fig. 2.7B), which are known to distinguish the INP, GMC/immature neuron, and mature neuron states, respectively, and found their UMAP positions matched well with their pseudotime alignments (Fig. 2.7A).

Table 2.1: Marker genes for scoring differentiation states

progenitors /INPs	GMCs	immature neurons/iN	mature neurons/mN	glia	astrocytes
CycE, mira, dpn	*insb, insc, spdo, nerfin-1, dap*	*Hey, E(spl)m6-BFM*	*nSyb, lncRNA:noe, jim*	*repo, gcm*	*Gat, alrm*

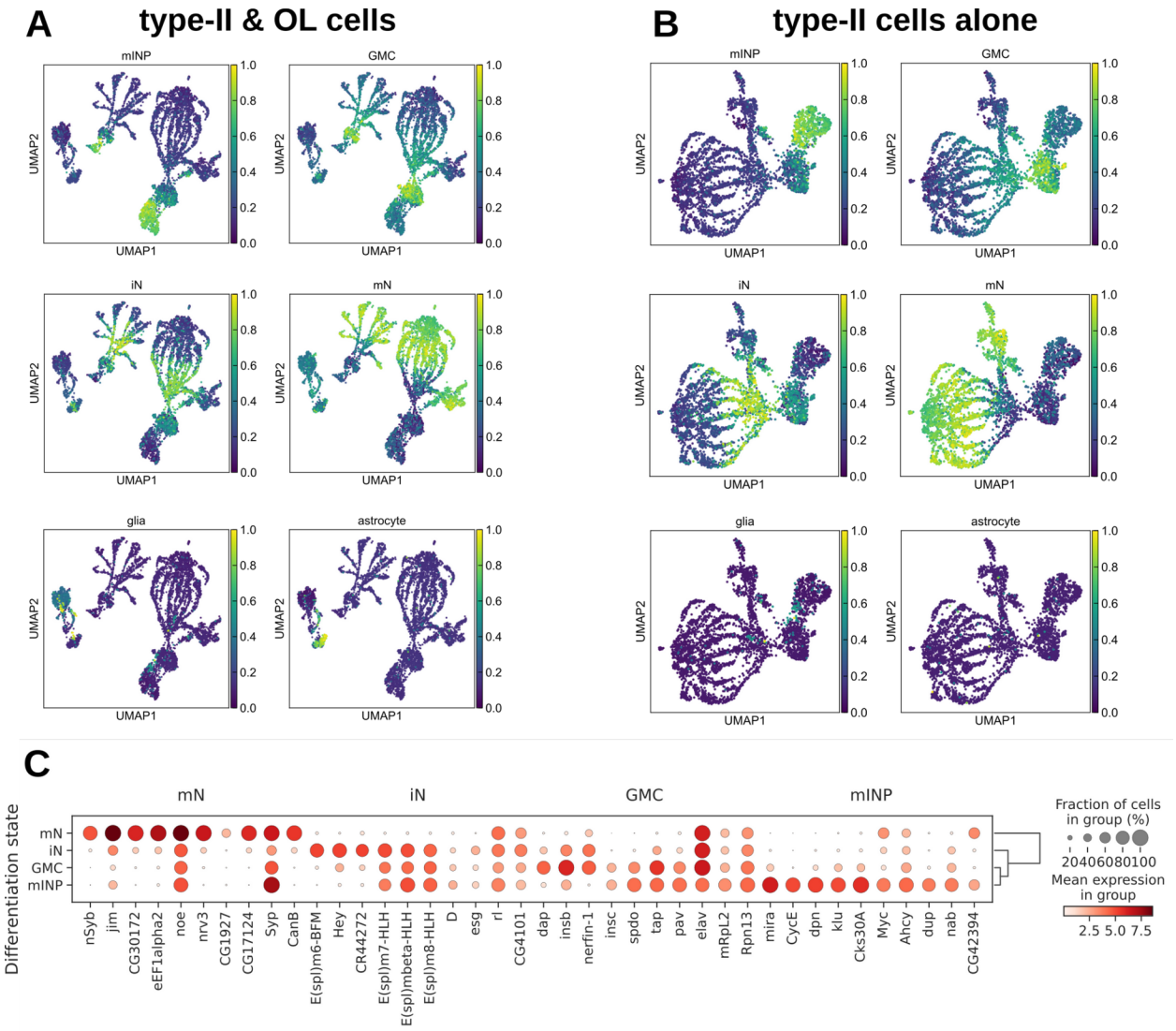

Fig. 2.6: Marker gene-based differentiation state scoring enables robust identification of cell differentiation state without manual annotation

(A) All cells were scored using the score_genes function from scanpy with the following marker genes defining each differentiation state (see table below). These scores were normalized to be within the range of [0,1], with 1 indicating that all of the marker genes for that cell type were expressed at high levels in that particular cell. (B) Type-II NB derived cells were scored as described in (A). (C) Marker gene analysis revealed genes that specify clusters of cells in distinct maturation/differentiation states. Many GMC marker genes are also highly expressed in INPs/progenitors. Although pseudotime analysis provides a more holistic view of a gene's dynamic change along the cell differentiation trajectory (Fig. 2.7), this small set of genes are robust identifiers for determining cell differentiation states. INP, intermediate progenitor cell; GMC, ganglion mother cell; iN, immature neuron; mN, matured/maturing neuron.

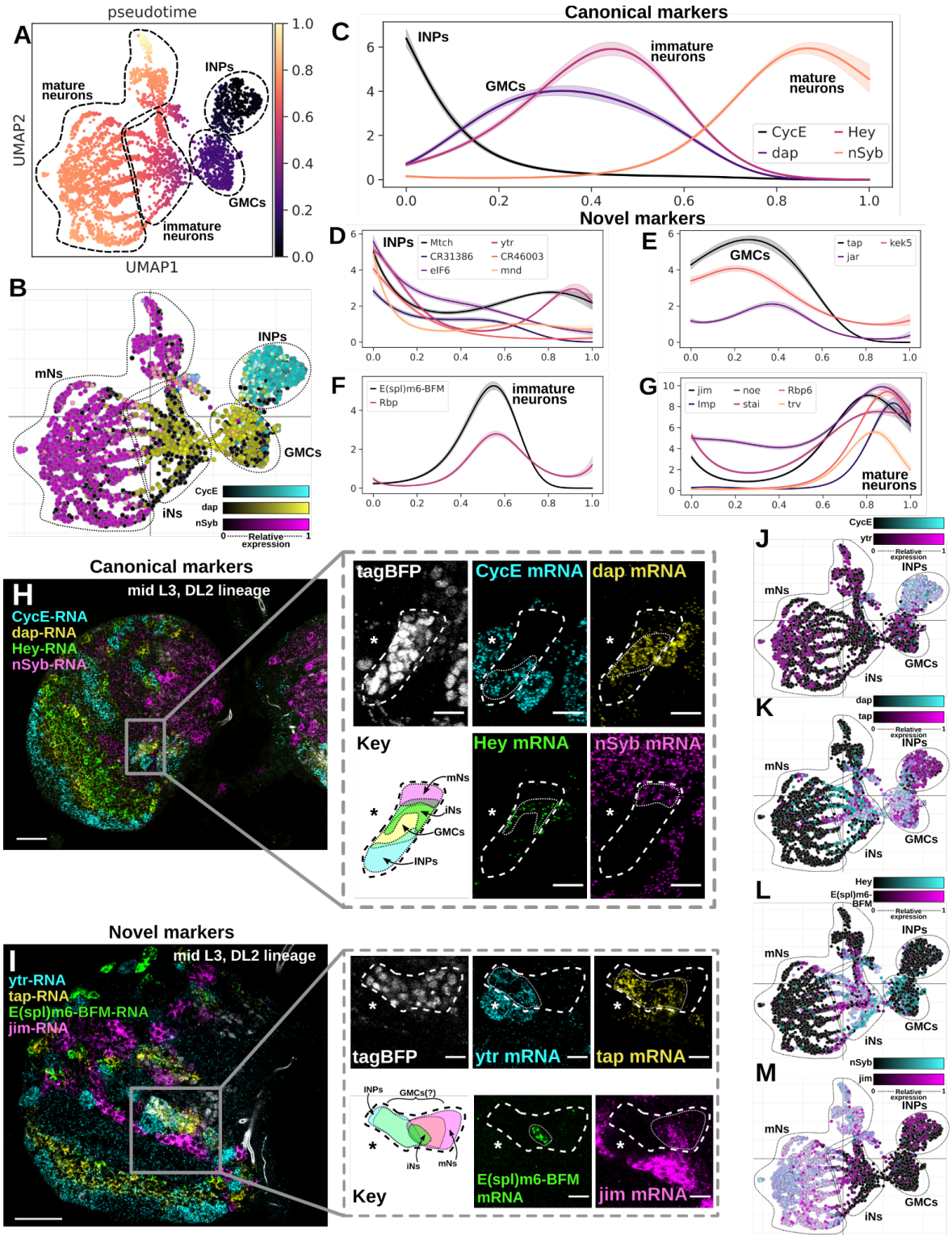

Fig. 2.7: Pseudotime analysis reveals signature genes that vary along the cell differentiation axis

(A) Pseudotime analysis establishes a global ordering of cells along the differentiation

state axis. (B) A multi-color UMAP expression plot generated by the MiCV web tool shows the expression of 3 canonical marker genes for the INP, GMC, and mature neuron states. (C) The pseudo-temporal expression pattern of 4 genes that are known markers for the 4 major differentiation states. (D-G) Pseudo-temporal expression patterns of groups of marker genes that do not have known functions associated with cellular differentiation state. These gene expression trends are similar to the known marker genes plotted in (C). (H, I) HCRv3 in situ mRNA staining images for both known (H) and novel (I) differentiation state marker genes in single z-slices of the DL2 lineage of mid 3rd instar larval brains. UAS-hH2B::2xtagBFP is driven under the control of R9D11-Gal4 and marks the type-II lineages. Asterisks denote the location of the putative type-II NB. Thick dashed lines denote the boundaries of the tagBFP labeled type-II NB progenies. Thin dotted lines denote the boundaries of type-II progeny cells expressing indicated mRNAs. (J-M) Multi-color UMAP expression plots illustrate the expression pattern of the canonical and novel marker genes from (H) and (I), respectively. Scale bars: 30 μm in overviews of (H, I), 10μm in insets of (H, I).

To describe the dynamics of gene expression across pseudotime, and thus the differentiation process, we fit a gene expression trend line to each gene detected in our scRNA-seq dataset using PyGAM (Servén et al., 2018). Indeed, we found that the expression peaks of four marker genes, i.e. CycE for INPs (Yang et al., 2017), dap for GMCs (Lane et al., 1996; de Nooij et al., 1996), Hey for a subset of the transient immature neuronal state (Monastirioti et al., 2010), and nSyb for maturing neurons (Deitcher et al., 1998), aligned in this exact differentiation order along the calculated pseudotimeline (Fig. 2.7C). Hence, we can use the relative expression levels of these genes to approximate the boundaries of the continuously changing differentiation states (Fig. 2.7A, dashed lines) in pseudotime. Subsequently, we performed gene expression trend clustering using phenograph (Levine et al., 2015) to screen novel putative marker genes whose expression trend matched one of the four known marker genes' (Fig. 2.7D-G). Independently, we used a marker gene-based differentiation state scoring (Wolf et al., 2018) strategy to separate these differentiation stages and found similar sets of marker genes (Fig. 2.6). Interestingly, many of the putative marker genes do not have any known function related to neural differentiation. Further pathway analysis and gene manipulation studies will be needed to explore their exact roles in type-II neurogenesis.

Nonetheless, we profiled the in situ expression patterns of some putative marker genes we identified in this analysis. We first synthesized HCRv3 probes against the canonical makers CycE, dap, Hey, and nSyb transcripts (Methods) and used these probes to investigate their expression pattern in the type-II NB derived cells using our novel reporter fly. As predicted, these genes form largely non-overlapping expression patterns in the larval brain (Fig. 2.7H, left panel). We found that CycE transcripts were expressed in large neuroblasts as indicated by the large cell bodies (Fig. 2.7H, right panels, asterisk) and in smaller tagBFP positive cells as a marker for replicating INPs. As predicted, dap, Hey, and nSyb transcripts expressed in bands of cells that were sequentially positioned away from the neuroblast (Fig. 2.7H, right panels, dashed lines). Next, from the gene expression trend clustering result (Fig. 2.7D-G), we selected four candidate markers and performed similar HCR in situ mRNA profiling. The in situ results suggest that ytr, tap, E(spl)m6-BFM, and jim transcripts express in unique patterns (Fig. 2.7I, right panels) and the co-expression MiCV plots indicate that these markers largely overlap the canonical makers in the respective cells (Fig. 2.7J-M). In particular, E(spl)m6-BFM, and jim were expressed almost exclusively in immature neurons and maturing neurons, respectively (Fig. 2.7L-M). However, while the putative INP marker ytr expressed in 96% of all the INPs, it also expressed in 37% of GMCs and 38% of maturing neurons (Fig. 2.7J). This observation indicates that ytr broadly expresses in INPs and that its expression may be selectively maintained in a subset of GMCs and their progeny neurons. The putative GMC marker tap appears to express in subsets of INPs and approximately half of the immature neurons (Fig. 2.7K). This suggests that tap may be a gene that defines one daughter neuron during their mother GMC's terminal cell division.

Though many genes that trend along the differentiation state axis are potentially interesting, we highlight here the gene E(spl)m6-BFM, a member of the Notch-responsive subgroup of the "enhancer of split" family of transcription factors (Lai et al., 2000). This family of proteins is responsible for regulating a variety of developmental

processes (Maier et al., 1993) and their group's function in balancing the self-renewal of differentiation in the type-II neuroblasts of Drosophila has recently been described (Li et al., 2017). However, the specific function or restricted spatial expression of E(spl)m6-BFM in the developing larval brain has not been established. Based on our analysis, E(spl)m6-BFM marks a subset of the cells in the transient immature neuronal state which comes about directly after the mother GMC's terminal cell division. This is similar to Hey, a previously identified immature neuron marker, which is upregulated in only one of the two daughter neurons of this terminal GMC division (Monastirioti et al., 2010) and activates in a Notch-dependent manner. Our scRNA-seq data indicates that E(spl)m6-BFM is expressed in both Hey+ cells and Hey- cells that have similar pseudotime values, though Hey+/E(spl)m6-BFM- cells are also present (Fig. 2.7L). To validate this, we used HCR probes for both Hey and E(spl)m6-BFM and identified subsets of immature neurons that were only Hey+, only E(spl)m6-BFM+, or Hey and E(spl)m6-BFM double-positive (Fig. 2.8). Similar to E(spl)m6-BFM, Rbp, a protein known to be functionally required for synaptic homeostasis and neurotransmitter release (Liu et al., 2011; Müller et al., 2015), is also upregulated only in this immature neuronal subset (data not shown). Further study will be desired to understand why either of these genes undergo a burst of expression in the immature neuronal state, and to establish their functional roles at the protein level.

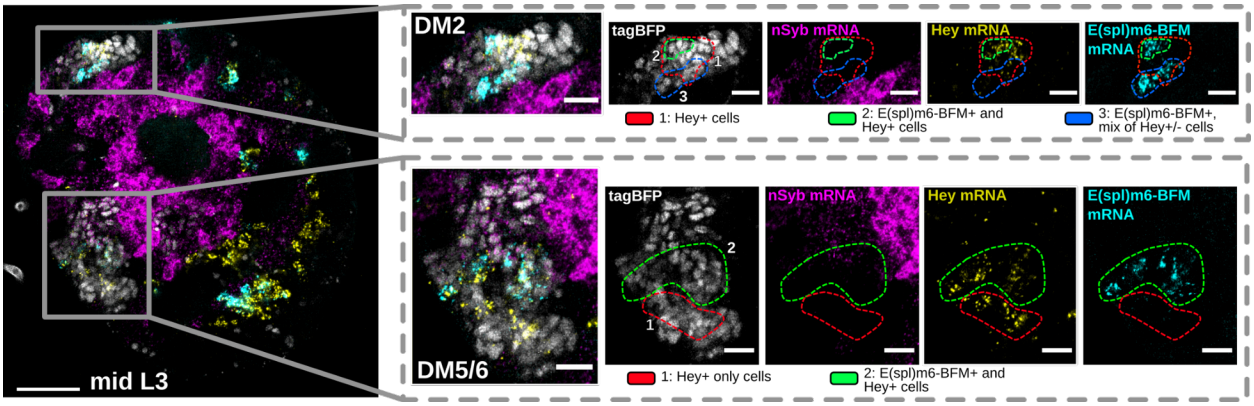

Fig. 2.8: The genes Hey and E(spl)m6-BFM mark an immature neural state
(A) Single z-slice of mid L3 larval brain; type-II lineages are marked with R9D11::hH2B-2xtagBFP. In lineages DM2 (top) and DM5/6 (bottom), Hey and E(spl)m6-BFM are clearly shown to be closely expressed in a small and restricted cell population that are

not nSyb+ maturing neurons. As previous studies revealed that Hey is strictly expressed in immature neurons, it is highly plausible to hypothesize that E(spl)m6-BFM is also expressed strictly in the same state, yet with a differential pattern compared to that of Hey. Scale bars: 30 μm in overview, 10μm in insets.

INP and GMC sub-clustering enables the identification of novel maturation pathways that are convolved with the canonical Dichaete, grainy-head, eyeless transitions

Having used pseudotime analysis to define the major differentiation states in the type-II neurogenesis process, we next characterized the cellular heterogeneity within these states using automated scRNA-seq clustering analysis. Such analysis may or may not obviously reflect previously established models of cell type differentiation/diversity, especially when this diversity could refer to any of/all the axes of cell type differentiation (Fig. 2.5A). Nonetheless, we performed Leiden clustering (Traag et al., 2019) with a low resolution (0.6) and overlaid the result on the reprojected UMAP (Fig. 2.9A, left). We found that cluster 1 and 0 included 561 and 563 cells, which correspond to the INP and GMC populations in the above-mentioned pseudotime analysis, respectively.

Subsequently, we took these putative INP and GMC cells and found they could be clustered into four groups of INPs and four groups of GMCs (Fig. 2.9A, right).
To discover which genes distinguished each subcluster, we performed logistic regression-based marker gene analysis and plotted the top 10 genes that defined the INP (Fig. 2.9B) or GMC subclusters (Fig. 2.9C). We found that this clustering result reflects a convolution of the lineage-specific canonical Dichaete, grainy-head, eyeless transitions outlined in (Bayraktar and Doe, 2013), which have been indicated to sequentially express in young to old INPs over the course of their division cycles (Bayraktar and Doe, 2013). D expression was rather specific in 74% of subcluster 4 INP cells and in 78% of subcluster 1 GMC cells, while only expressing in fewer than 28% of other subcluster cells (Fig. 2.9D). On the contrary, grh and ey expressions are intermingled in the other subclusters (Fig. 2.9E and F, respectively).

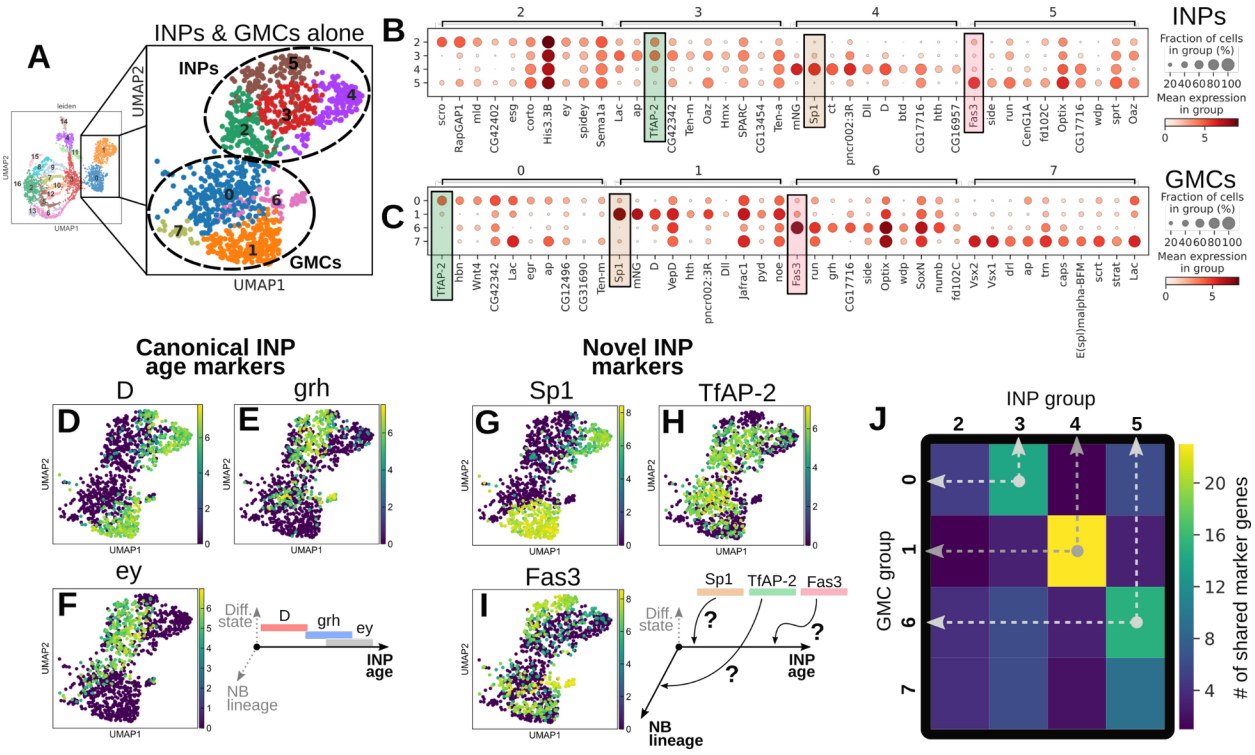

Fig. 2.9: Sub-clustering of INPs and GMCs reveals transcription factors beyond the canonical D-grh-ey transition that vary along a combination of the NB lineage and INP division number patterning axes

(A) Left panel, leiden clustering reveals INPs and GMCs to be in cluster 1 and 0, respectively. Right panel, higher resolution clustering on separated INPs and GMCs further divides them into 4 subclusters each. (B, C) Marker gene analysis revealed that mostly transcription factors specific INP and GMC subclusters, respectively. (D-F) Expression UMAP plots of the well-established temporally-varying INP genes D, grh, and ey, respectively. D appears to separate cleanly at the mRNA level in the INPs of our dataset, however, grh and ey are broadly co-expressed. (G-I) Expression UMAP plots of the INP/GMC cluster-specific genes Sp1, TfAP-2, and Fas3, which are found to correlate INP subclusters 3, 4, and 5 to GMC subclusters 0, 1, and 6, respectively. (J) A correlation plot shows the number of top 100 marker genes that are shared between each INP and GMC subcluster. This simple similarity metric indicates a hypothesis that cells in GMC subclusters 0, 1, and 6 are the direct progenies of cells in INP subcluster 3, 4, and 5, respectively. INP group 2 and GMC group 7 are both clearly distinct from the other INP and GMC subtypes, but share very few marker genes and so are unlikely to be related.

Interestingly, we found that Sp1, TfAP-2, and Fas3, among the top marker genes in this

clustering analysis, not only expressed in segregated subclusters, but also marked both INP and GMC subclusters (Fig. 2.9G, 3H, and 3I, respectively). We suspected that the GMC subclusters specified by these genes might be the direct progenies of the INP subclusters that carry over the Sp1, TfAP-2, and Fas3 transcripts. We subsequently counted the number of top 100 marker genes that were shared between each of the INP and GMC subclusters. The correlation plot strongly suggests that GMC subclusters 0, 1, and 6 are likely the progeny of INP subclusters 3, 4, and 5, respectively (Fig. 2.9J).

The choice of clustering resolution can be somewhat arbitrary, and the 8 subclusters for INPs and for GMCs here may represent a surface level of INP patterning that can be further broken down into more subtypes. Since we saw a clear link between 6 of these 8 subclusters, we decided to pursue in situ validation experiments for the marker genes identified at the 8-subcluster resolution in follow up experiments, and aimed to do so in an exploratory manner, taking Sp1, TfAP-2, and Fas3 (the top marker genes for the relevant GMC subclusters) as promising marker genes to investigate further.

The transcription factor *Sp1* is expressed in young INPs throughout the DM1-6 and DL1 lineages and marks a unique neural progeny

We first aimed to in situ profile the transcript expression of Sp1, a Cys2His2-type zinc finger transcription factor that is necessary for the specification of type-II neuroblasts (Álvarez and Díaz-Benjumea, 2018). We reasoned that this, along with the apparent coexpression of Sp1 with D in the INPs of our scRNA-seq dataset (Fig. 2.9D, G, respectively), would imply that Sp1 may be broadly expressed in young, newly-matured INPs of most type-II NB lineages. We synthesized HCRv3 probes against Sp1 and D transcripts (Methods) and validated their specificity using gene-trap reporter flies (Fig. 2.10).

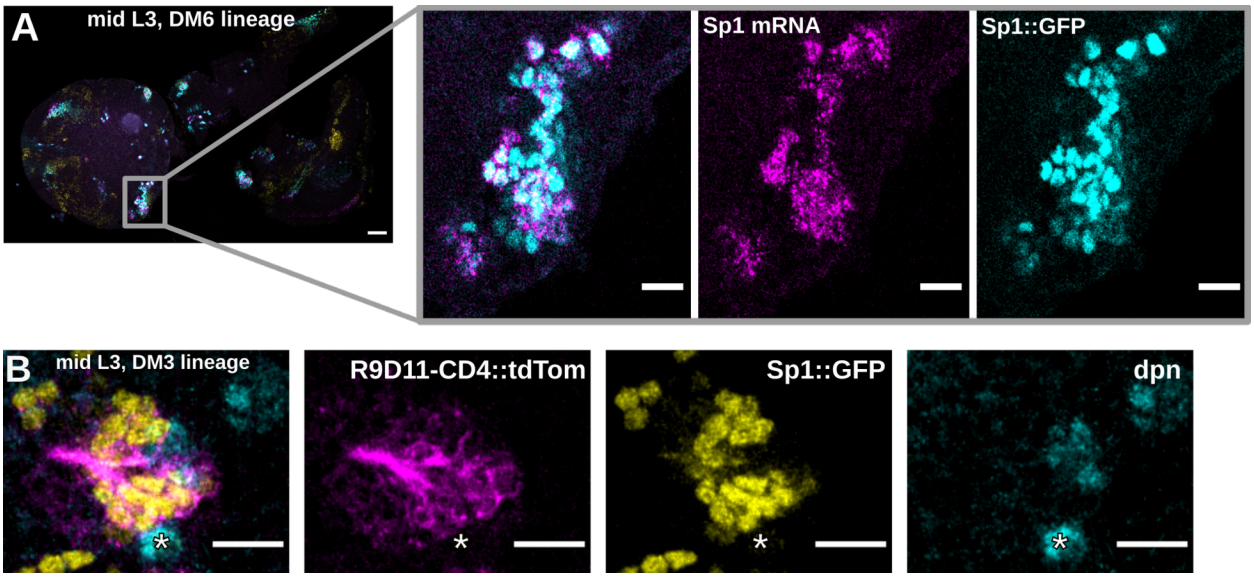

Fig. 2.10: Sp1::GFP fusion protein and Sp1 mRNA co-localize in situ and label both dpn+ INPs and axon-producing neurons
(A) Single z-slice of a mid 3rd instar Sp1::GFP transgenic larval brain, showing native fluorescence of GFP in cyan (right), HCR stained Sp1 mRNA in magenta (middle), and a composite of the two (left). Sp1 mRNA signal is made up of puncta scattered around the labeled cells. Sp1 protein, being a transcription factor, leads to GFP expression being largely localized to the nucleus. (B) Single z-slice of a mid 3rd instar ;R9D11-CD4::tdTomato;Sp1::GFP larval brain stained with an antibody specific to dpn. The co-localization of dpn, Sp1, and membrane-bound tdTomato proteins indicates that Sp1 is translated in both neurons and INPs of the type-II lineage, as evidenced by the labelling of cells that either produce membrane-tdTomato-labeled axons or are dpn+, respectively. Scale bars: 30 μm in the overview of (A); 10 μm in insets of (A) and (B).

When assessing their expression patterns in the type-II NB derived cells, we found that Sp1 mRNA was expressed prominently in all type-II lineages with the possible exception of DL2 (Fig. 2.11A, B). On the contrary, D mRNA expressed prominently in DM1-3, and in much smaller subsets of cells in lineages DM4-6 (data not shown), which is consistent with previous observations (Bayraktar and Doe, 2013).

Our scRNA-seq data indicates that while Sp1 co-expressed with D in more than 30% of INPs (Fig. 2.11C), 8% and 16% of all INPs are Sp1+/D- and Sp1-/D+, respectively. To validate the presence of these INP populations in situ, we used our HCR protocol to co-stain Sp1 and D mRNA (Fig. 2.11D). We show that, for instance in the DM6 lineage, an

Sp1+, D+ INP progeny can be identified directly adjacent to cells where either Sp1 or D is exclusively expressed (Fig. 2.11D, enlarged box). Furthermore, we overlaid Sp1 or D expressions on the UMAP plot, and found that these two transcripts continue to express in maturing neurons of two exclusive subsets (detailed below). This is consistent with a previous study, which found that the D expressing young INPs specifically give rise to D expressing neurons (Bayraktar and Doe, 2013). Therefore, we hypothesize that Sp1+/D+ INPs may transition to Sp1 or D exclusively expressing INPs, which give rise to distinct neural subtypes. To specify whether Sp1 protein is expressed in neurons, we labeled the type-II progenies with a membrane-bound tdTomato (R9D11-CD4::tdTomato) to visualize neuron's characteristic axonal projections and coupled with an Sp1::GFP reporter line. We show as an example that the DM3 lineage generates many neurons that form a tdTomato+ neurite bundle that are also GFP+, which indicates the generation of Sp1+ neural progeny (Fig. 2.10B).

Next, we wondered whether Sp1 is like D that expresses strictly in young INPs. We quantified our scRNA-seq data and found that Sp1 coexpressed with the two canonical late INP markers grh and ey only in a small subset of INPs (Fig. 2.11C). Taken together, these data support the hypothesis that Sp1, much like D, is expressed broadly in INPs with low division numbers and that these INPs are responsible for producing a neural progeny similarly marked by Sp1 expression that is distinct from the D+ neural progeny.

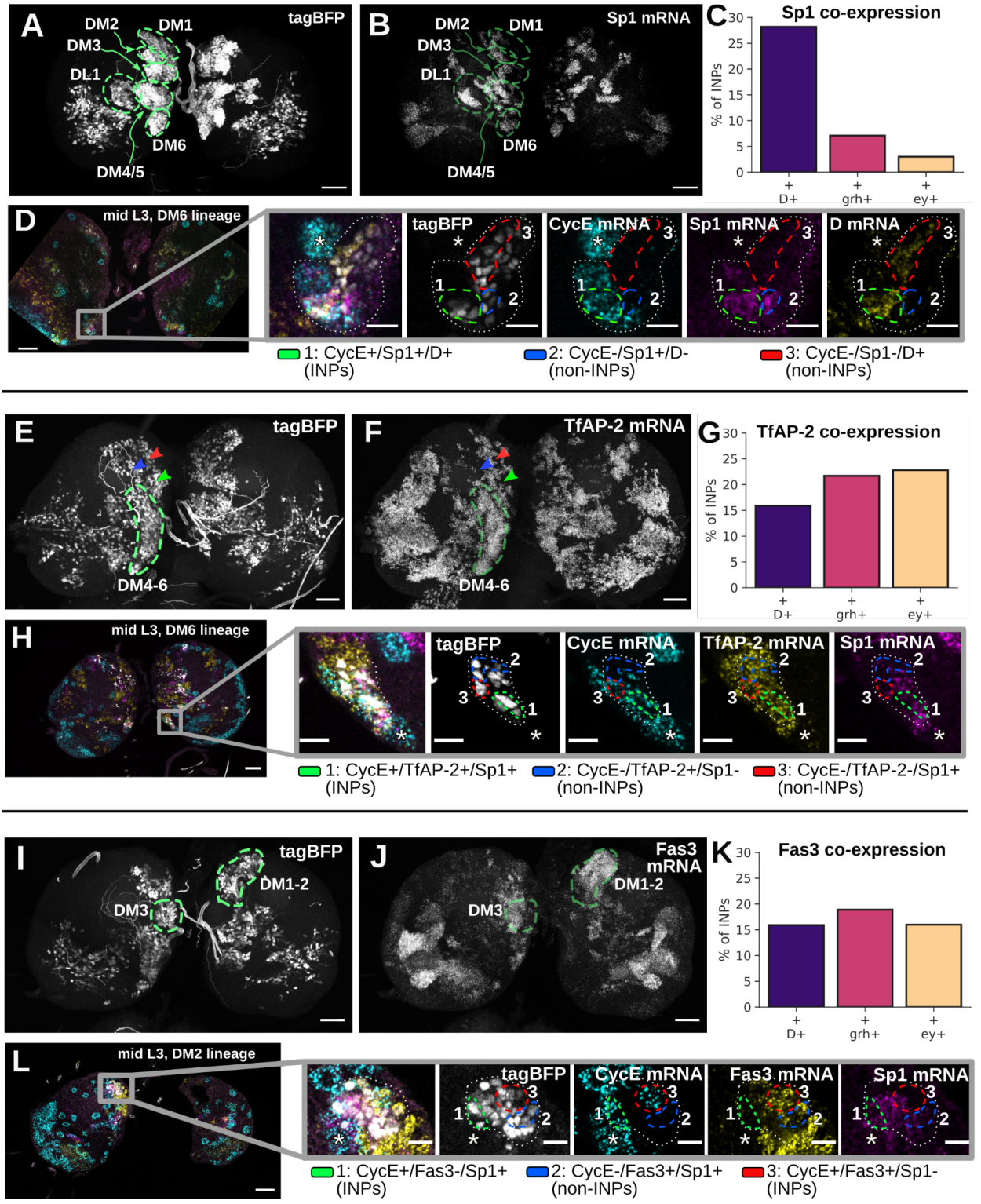

Fig. 2.11: Sp1, TfAP-2, and Fas3 are each expressed by INPs of specific NB lineages

(A, B) Maximum Z-projections (45μm thick) show tagBFP fluorescence and Sp1 mRNA

HCR staining in an L3 larval ;;R9D11-Gal4/UAS-H2B::tagBFP fly brain, respectively. Green dashed lines indicate the expression of Sp1 mRNA in all type-II NB derived lineages except for DL2. (C) Co-expression quantification of Sp1 with D, grh, and ey in all INPs (n=561). (D) HCR staining showcases the expression patterns of Sp1 and D mRNAs in lineage DM6. Dashed lines highlight region 1 of INPs that co-express Sp1 and D mRNA, region 2 of non-INP cells where Sp1 mRNA alone is detected, and region 3 of non-INP cells where D mRNA alone is detected. White dotted lines denote the DM6 lineage boundary. Asterisks denote the position of the DM6 neuroblast. (E, F) Maximum Z-projections (45μm thick) as in (B,C), with TfAP-2 mRNA HCR staining. Within the type-II NB lineages, TfAP-2 mRNA is highly expressed in cells belonging to DM 4-6 (dashed lines) and possibly DL1. Though some expression is seen nearby to DM1-3, TfAP-2 is not expressed in tagBFP+ cells belonging to those lineages (arrowheads). (G) Co-expression quantification of TfAP-2 as in (C). (H) HCR staining showcases the expression patterns of CycE, Sp1 and TfAP-2 mRNAs in lineage DM6, where we can find TfAP-2 expressed in CycE+ INPs that have Sp1 expression (green dashed lines) or not (red dashed lines), as well as in CycE- progeny cells (blue dashed lines). (I, J) Maximum Z-projections (45μm thick) as in (B,C), with Fas3 mRNA HCR staining. Within the type-II NB lineages, Fas3 mRNA is highly expressed in cells belonging to DM 1-3 (dashed lines). (K) Co-expression quantification of Fas3 as in (C). (L) HCR staining showcases the expression patterns of CycE, Fas3, and Sp1 mRNAs in lineage DM2. We find a clear expression of Fas3 in both INPs and their progeny in NB lineage DM2, where we can find Fas3 expressed in CycE+ INPs that have Sp1 expression (green dashed lines) or not (red dashed lines), as well as in CycE- progeny cells (blue dashed lines). Scale bars: 30 μm in (A, B, E, F, I, J) and in overviews of (D, H, L); 10 μm in insets of (D, H, L). Minimum expression threshold: ln(CPM+1)>4.5 in (C, G, K).

The transcription factor TfAP-2 and cell adhesion molecule Fas3 are each expressed in INPs of specific type-II neuroblast lineages

We next characterized the spatial expression patterns of TfAP-2 and Fas3, selected markers for the other two major putative INP subtypes identified in our low-resolution clustering (Fig. 2.9). We generated HCR probes against mRNA of TfAP-2 and Fas3 in a similar manner to Sp1 and probed their expression in reporter flies in order to identify which type-II NBs generate their respective INP subsets. Unlike Sp1, however, TfAP-2 and Fas3 transcripts are expressed much more broadly across the brain and are not restricted to the type-II lineages (Fig. 2.11F, J).

Within the type-II progenies, TfAP-2 mRNA appeared to be expressed prominently in INPs of the DM4-6 lineages as well as a subset of their downstream progeny (Fig. 2.11E, F, green outline). However, we did not observe strong TfAP-2 expression in any other lineage, implying that expression of this marker is primarily lineage restricted (Fig. 2.11E, F, arrowheads). Interestingly, TfAP-2 co-expressed in fewer D+ but many more grh/ey+ INPs than Sp1 does in our scRNAseq data, which indicates that TfAP-2+ INPs have likely undergone some cell divisions before expressing this marker gene (Fig. 2.11C vs 4G). Although TfAP-2 expresses in fewer lineages than Sp1, our scRNA-seq data (not shown) and in situ profiling (Fig. 2.11H) showed that these two genes do indeed co-express in cells belonging to those few lineages. TfAP-2 plays broad roles in development (Monge et al., 2001), but in the context of the central brain it has been shown to play a role in developing and maintaining the neural circuitry required for night-sleep in adult flies (Kucherenko et al., 2016). Consistently, we found in our scRNAseq data that TfAP-2 expressed in a subset of neurons that are distinct from the Sp1+ or D+ population (data not shown). TfAP-2's expression in neurons is distinct from the previously identified late INP progeny genes grh and ey; the latter two were not found in neurons in our scRNA-seq data (data not shown). TfAP-2 (ap-2) is significantly orthologous to the human transcription factors TFAP2A/B (Flybase curators, 2019), and its role in sleep can be traced back to C. elegans (Turek et al., 2013). Taken together, this would imply that at least this particular role for TfAP-2 in the central brain may be evolutionarily conserved and that the neurons generated by TfAP-2+ INPs in the DM4-6 lineages may play a role in night-sleep circuit maintenance.

Based on our in situ RNA staining, Fas3 mRNA was found to express most prominently in the INPs of DM1-3 (Fig. 2.11I, J). Similar to TfAP-2, our scRNAseq data suggests that Fas3 co-expressed in fewer D+ but many more grh/ey+ INPs than Sp1 does, which indicates that Fas3 INPs have likely undergone some cell divisions before expressing this marker gene (Fig. 2.11C vs K). Again, our scRNA-seq data (not shown) and in situ profiling (Fig. 2.11L) showed that Fas3 and Sp1 co-express in a significant fraction of cells. Fas3 is interesting as a marker gene for INPs as it is not a transcription factor but

rather a membrane-bound, homophilic cell adhesion molecule that plays a strong role in synaptic targeting and axonal guidance in a subset of neurons in the central and peripheral nervous systems (Kose et al., 1997; Snow et al., 1989), along with cell adhesion-mediated morphological development throughout the entirety of the fly (Wells et al., 2013). Why Fas3 would be expressed so strongly in a subset of INPs is unknown.

A unique combination of transcription factors and surface molecules define putative neural sub-progenies of young INPs

With low resolution (0.6) global clustering, our scRNA-seq data already showed a much greater subtype diversity in neurons (12 clusters) than in GMCs or INPs (1 cluster each) (Fig. 2.12A). We performed logistic-regression based marker gene analysis on these specific clusters to identify the top 100 marker genes for each cluster that are most uniquely expressed with the top 10 marker genes of clusters 4, 6, and 8 are plotted in Fig. 2.12B (full plot for all clusters are shown in Fig. 2.13). Subsequently, we analyzed the top 100 marker genes using the DAVID Functional Annotation Tool (Huang et al., 2009a, 2009b) in order to identify sets of genes that form functionally associated groups based on associated gene ontology (GO) terms. We identified the first GO term from the top three highly enriched functional groups and find that these terms indicate that transcription factors and surface molecules are predominant markers for these three (Fig. 2.12C), as well as all other neural subsets (data not shown).

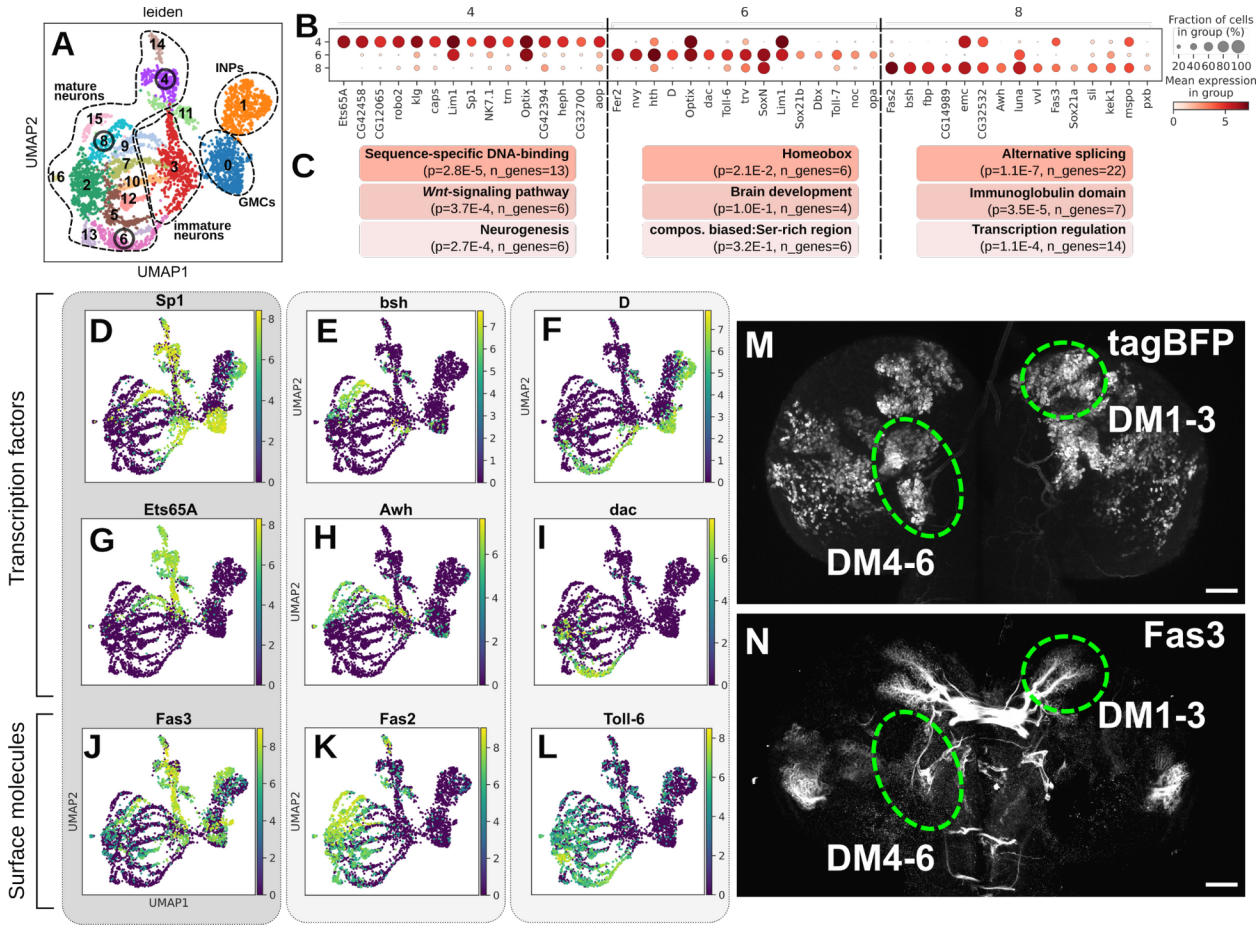

Fig. 2.12: A unique combination of transcription factors and surface molecules define putative neural sub-progenies of young INPs

(A) Automatic Leiden clustering (resolution = 0.6) of the type-II scRNA-seq data, with putative neural subtypes 4, 6, and 8 outlined, representing the Sp1, bsh, and D+ neural progenies, respectively. (B) Marker gene detection for the three selected neural subtypes showing the top 15 marker genes as identified using the t-test_overestim_var function in scanpy. (C) Top gene ontology (GO) functional annotations for the top 100 marker genes for cells in each of clusters 4, 6, and 8, respectively (p-values are Benjamini corrected; n_genes refers to the number of marker genes annotated with the respective GO term). (D, E, F) Log-fold expression values of Sp1, bsh, and D, respectively, showing three unique neural lineages are marked by these three transcription factors. (G, J) Log-fold expression values of the transcription factor Ets65A and the cell surface molecule Fas3 that mark the Sp1+ neural progeny. (H, K) Log-fold expression of the transcription factor Awh and surface molecule Fas2 that mark the bsh+ neural progeny. (I, L) Log-fold expression of the transcription factor dac and surface molecule Toll-6 that mark the D+ neural progeny. (M, N) Maximum Z-projections show tagBFP fluorescence and Fas3 antibody staining in an L3 larval ;;R9D11-Gal4/UAS-H2B::tagBFP reporter fly brain, respectively. It appears that neurons

from DM1-3 that produce commissure-crossing axons are prominently labeled by Fas3, whereas neurons from DM4-6 are largely unstained. Scale bars: 30 μm in (M, N).

In the (Bayraktar and Doe, 2013) study, bsh was found to express in a non-overlapping subset of neurons that do not express D in the young INP progeny. The same study also specified that there are other young INP derived neurons are Bsh- and D-, whose markers were not identified using the available method and antibody probes. Interestingly, our automatic analysis reveals that neuron clusters 4, 6, and 8 differentially express Sp1+, D+, and bsh+ respectively (Fig. 2.12B). As we show that Sp1 expresses in young INPs, it is plausible that Sp1 is a marker gene that labels the previously unspecified young INP derived neurons. Indeed, our scRNAseq data shows that Sp1, D, and bsh were expressed in three distinct maturing neuron populations (Fig. 2.12D, F, and E, respectively). This in silico analysis permits rapid identification of transcription factors that potentially belong to the same regulatory pathways to specify neuronal fate. For example, selected from the specific marker gene list, the transcription factors Ets65A, dac, and Awh are highly co-expressed with neurons expression Sp1, D, and bsh, respectively (Fig. 2.12G, I, H, respectively).

Distinct surface molecules are also differentially expressed in different subsets of neurons, which may indicate their roles in forming functionally distinct circuits (Fig. 2.12J-L). Among them, Fas3 appears to co-express in a large proportion of Sp1 neurons, regardless of their low degree of co-expression in the INP and GMC stages (compare Fig. 2.12D vs J). To validate that Fas3 protein is translated in neurons of this developmental stage, we used a Fas3 antibody to stain our novel type-II lineage reporter fly and found that it labels neurons in the DM1-3 lineages that form neurite bundles across the commissure (Fig. 2.12M, N). It is plausible that the expression of Fas3 in INP may play a role in enabling some of the neural progenies of DM1-3 to either form these axonal bundles or for them to find their final targets across the commissure early on in the neural maturation process.

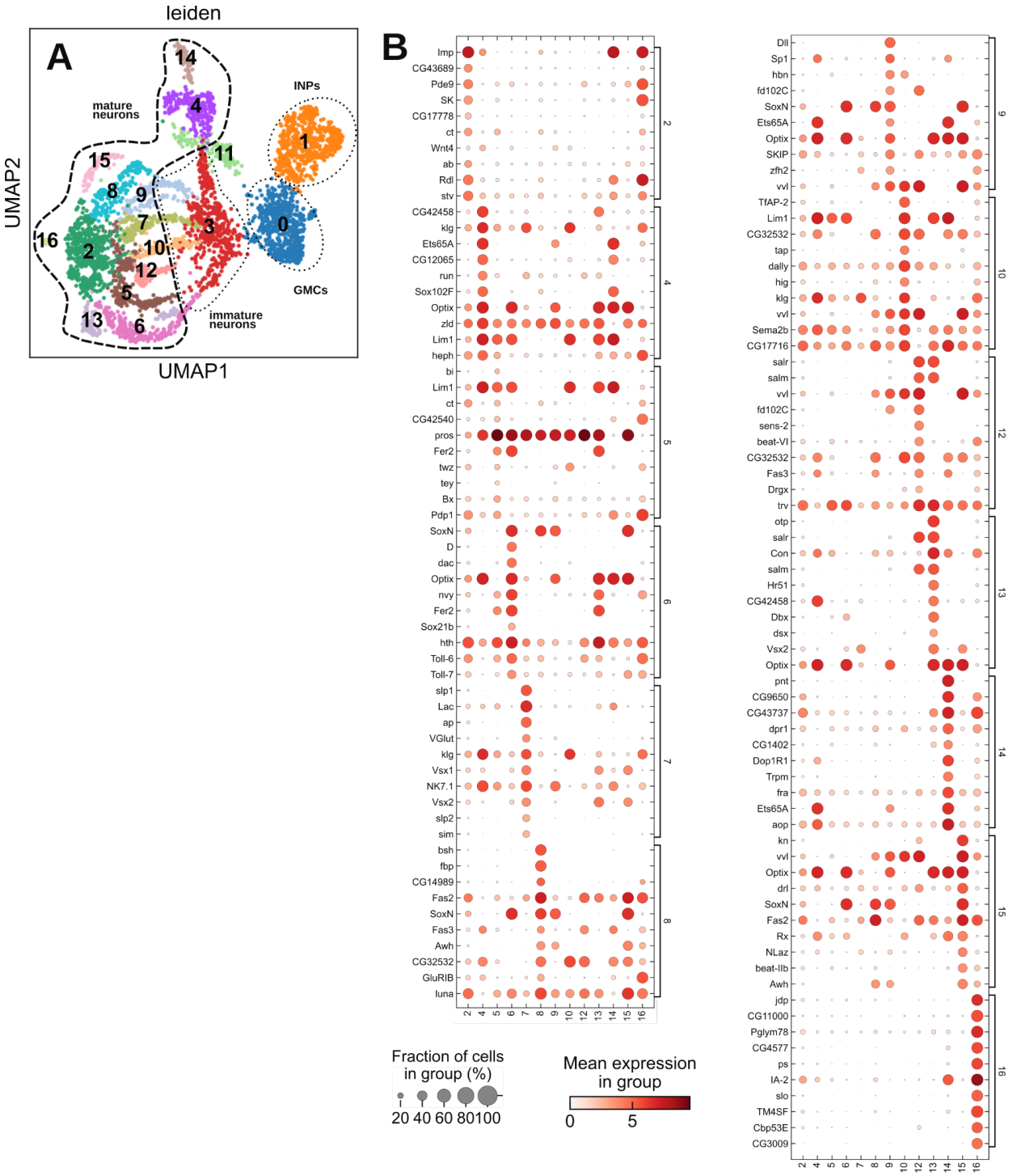

Fig. 2.13: Marker genes of subtype-specific immature/maturing neurons
(A) UMAP plot with automatic cluster assignments (resolution = 0.6). (B) The top 10 marker genes identified for each of the neural clusters in this scRNA-seq dataset.

Drosophila type-II neural lineages as a model system to study complex neurogenesis processes

To enable the brain's complex functions, vastly diverse neuronal types need to be rapidly generated at a very large scale during development. To reveal how neural stem cells populate the developing brain, efforts have been made to identify cell types and their lineage relationships. For instance, focuses on neurodevelopment in mouse (Habib et al., 2017; Han et al., 2018; Ponti et al., 2013; Saunders et al., 2018; Soldatov et al., 2019), human brain tissues (Habib et al., 2017) and the developing human prefrontal cortex (Zhong et al., 2018) revealed intermediate stem cells (and critical genes involved) as an important mechanism for rapid cortical expansion. Underlying this rapid and diverse differentiation process is the constant change of gene expression profiles in all cells. However, the molecular mechanisms that lead to functionally distinct neurons in the mammalian brain remain challenging to describe in detail. This is because, on the one hand, neuronal fate determination involves many genes, and on the other hand, neural progeny cells originating from distinct lineages undergo rapid migration, which leads to their intermingling nature in space.

Although they are the minority (8 stem cells per hemisphere) in the Drosophila central brain, the Drosophila type-II neural lineage has a neurogenesis process analogous to the mammal's rapid cortical expansion (Homem and Knoblich, 2012). Compared to their mammalian counterparts, the Drosophila type-II neural lineage has the advantage of being non-migrating in the larval stage. With proper labeling, type-II progeny cells of the same lineage can be identified as a segregated cell cluster. Importantly, the cells' spatial relationship within a cluster serves as a considering factor when determining the age and maturation stage of these cells (Boone and Doe, 2008; Homem and Knoblich, 2012). The small stem cell pool and mammal-like lineage composition make the Drosophila type-II neural lineage an attractive model to study the complex brain development process. In addition, many important genes and signaling pathways are conserved throughout evolution (Homem and Knoblich, 2012; Mariano et al., 2020; Ogawa and Vallender, 2014), which makes revealing the molecular mechanisms of

Drosophila type-II neuron differentiation a meaningful primer to study the human analogs in the outer subventricular zone.

Summary of this work

In this work, we used targeted single cell transcriptome analysis to advance our understanding of the Drosophila type-II neuron differentiation process. After initially separating the transcriptomes of the type-II neuroblast derived cells from those labeled in the optic lobes, we show that pseudotime analysis techniques can be used to define a maturation axis and extract putative marker genes that specify the INP, GMC, immature neuron and mature neuron differentiation stages. Broadly expressed, not limited to the type-II NB progenies, these marker genes of different maturation stages indeed form intersectional patterns that represent the spatial organization of the neurogenesis progress in the larval brain. Compared to previous antibody-based and gene manipulation-based screening strategies, scRNA-seq data permits a high-throughput assessment of the whole gene expression profile to rapidly identify candidate genes for functional study. For instance, in the past, Hey has been shown to mark one of the two immature neurons derived from the final cell division, and its role is exclusive as an inhibitor of Notch signaling in this immature neuron (Monastirioti et al., 2010). From our scRNAseq analysis, E(spl)m6-BFM, a member of the enhancer-of-split family of transcription factors (Lai et al., 2000), and Rbp, a rim-binding protein responsible for synaptic homeostasis and neurotransmitter release (Liu et al., 2011; Müller et al., 2015) are exclusively up-regulated in only the transient immature neuronal differentiation state directly after GMC division. These two marker genes can be used to guide the exploration of Hey- immature neurons in future studies. Functional knock-outs of these two genes will be critical to understanding their function in newly-born neurons as it pertains to their maturation and any early functional role they may play in the developing brain.

Further higher-resolution clustering of the INP and GMC cells identified transcriptomically correlated subclusters between these two stages, which supports the

idea that parallel maturation transitions happen at the same developmental time point. However, scRNA-seq data alone cannot distinguish whether these parallel transitions are due to the co-existence of earlier and newly born INPs in all NB lineages or due to the intrinsic differences among NB lineages. We therefore in situ profiled the marker genes selected from the scRNA-seq selected candidates and restored their missing spatial information that indicates the maturation stage as well as the NB lineage identity. In addition, combined with prior knowledge, whether a marker gene is expressed in younger or earlier born INPs can also be speculated. Our findings conclude that Sp1 is expressed in the young INPs of nearly all NB lineages, whereas TfAP-2 and Fas3 express in older INPs belonging to specific NB lineages. Interestingly, we found that Sp1 and TfAP-2 expressed not only in neural progenitors but also in maturing neurons. These transcription factors seem to intermingle with the NB lineage-specific D/grh/ey cascades in the INP stage, but eventually differentiate into completely exclusive neuron populations. Finally, higher-resolution clustering of neurons in our scRNAseq dataset revealed that transcription factors and surface molecules are predominant markers for distinct neuronal subtypes at the 3rd instar larval stage. This implies that most neurons of the type-II NB progenies have not started to gain their differentiated functions at this stage of development.

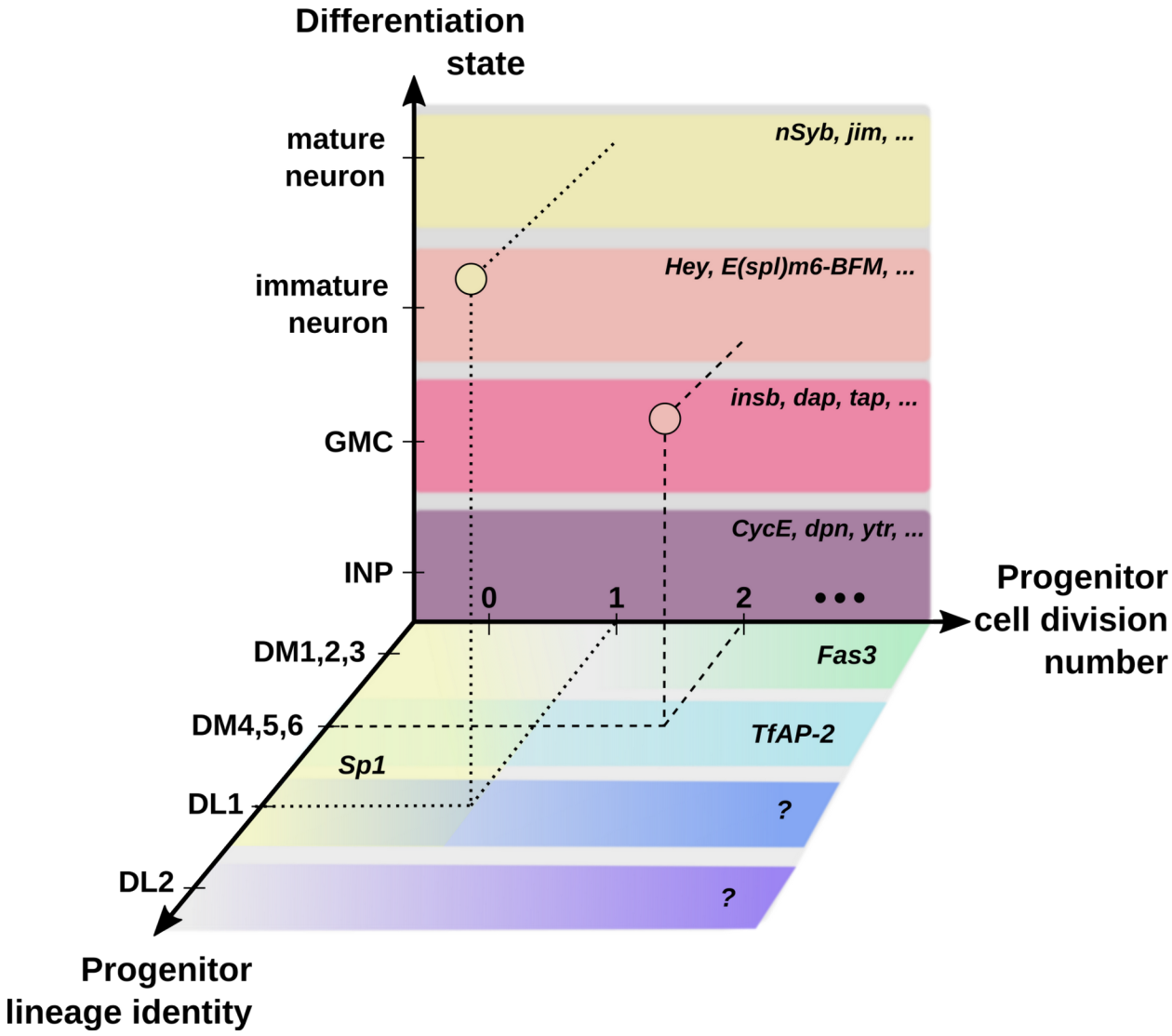

Fig. 2.14: A Drosophila type-II neuronal fate specification model illustrates the complex molecular network that determines the neural differentiation process
Despite its small scale and apparent simplicity, the complex interplay of molecular factors that vary along the differentiation state, lineage identity, and progenitor cell division number axes are responsible for determining the fate of each cell derived from the type-II neuroblasts of Drosophila. In this diagram, some of the most prominent molecular factors from the literature or identified and validated in this work are shown to occupy different domains along these three axes. Multi-time-point analysis and in situ validation will enable us to continue to fill in the blanks and develop a more complete roadmap of the type-II neurogenesis process across development.

Combining in silico scRNA-seq analysis and in situ mRNA imaging, we discovered many transcription factors and surface molecules that potentially play important roles in

generating neuronal subtypes in an NB-specific, INP-specific, or function-specific manner. These discoveries helped us to gain a comprehensive understanding of the molecular landscape along all three major neural developmental axes that define a cell's progenitor lineage identity, progenitor cell division number, and differentiation state (Fig. 2.14). This model provides a general guidance for biologists to disentangle the differentiation process in complex systems beyond the Drosophila brain.

Challenges and opportunities

We sequenced approximately 4000 cells that were neurons originating from 8 Drosophila type-II neuroblast lineages (16, if we assume no symmetry across the two central brain lobes). With low-resolution clustering, we identified 13 molecularly distinct neural subtypes. Increasing the clustering resolution just a bit higher we could identify more than 20 that are still distinct (data not shown). Similarly, as we show with the INPs/GMCs in our dataset, a low-resolution clustering can often mask the cellular diversity that is present in the system. As we know that each type-II neuroblast generates approximately 38 INPs throughout their developmental lifespan (Bayraktar et al., 2010; Bello et al., 2008), the presented clustering in this paper only captures part of the INP diversity. One straightforward thought is to increase the number of sequenced single cells so that higher clustering resolution may eventually reveal even the most subtle differences between each of the hundreds of INPs in the type-II system. However, as transcription factor cascades involved in INP division/maturation intertwine with those involved in NB specification and differentiation, we expect that the INP heterogeneity can be untangled somewhat using a higher clustering resolution but still fails to provide us with a coherent view of the complex lineage, maturation, and differentiation landscape we are attempting to characterize. These issues highlight the challenge of deconvoluting the INP maturation, NB lineage, and differentiation state axes and the need for a holistic, integrated approach to experimental design and subsequent bioinformatic analysis.

The data we have presented here were collected at a single developmental time-point

(late third instar), but we know that type-II neurogenesis precedes and continues after this stage. Repeating these scRNA-seq experiments at more developmental time-points will reveal more in what order molecularly-defined neural subsets are generated. Using recently developed analytical techniques to “stitch” these multi-time-point datasets together (Lin et al., 2019; Tran and Bader, 2019) will be advantageous to align all the cells along a unified developmental time axis. To overcome the limitation of the R9D11-Gal4 driver, which does not label neuroblasts nor the fully mature neurons, a permanent labeling strategy, similar to the one used in (Bayraktar et al., 2010) but covering all lineages more reliably for FACS, is required. More critically, such permanent labeling needs to be paired with technologies that provide single-lineage specification resolution, such as the introduction of single-neuroblast lineage barcoding techniques. Genetic constructs based around CRISPR-Cas9 (Raj et al., 2017; Spanjaard et al., 2018) and the Cre/Lox system (Kalhor et al., 2018; Pei et al., 2017; Weber et al., 2016) have been developed for this purpose, although which exact lineage was labeled by a particular barcode was still unknown. The introduction of a spectrally unique barcode for each neuroblast lineage, in a similar vein to the recently developed Bitbow lineage tracking strategy (Li et al., 2020; Veling et al., 2019), would be advantageous as they can provide direct in situ evidence for neuroblast lineage identity.

Finally, our work identifies several transcription factors that are specifically expressed in subsets of cells of the type-II neuroblast progenies. Our in silico and in situ results showed that their expressions are either constrained to particular developmental stages, or in subsets of cells that are born in different orders. It would be desired to perform follow up experiments to reveal whether these transcription factors play important roles in specifying the terminal fates of type-II neuronal subtypes.

References

Álvarez, J.-A., and Díaz-Benjumea, F.J. (2018). Origin and specification of type II neuroblasts in the Drosophila embryo. Development 145.

Bayraktar, O.A., and Doe, C.Q. (2013). Combinatorial temporal patterning in progenitors expands neural diversity. Nature 498, 449–455.

Bayraktar, O.A., Boone, J.Q., Drummond, M.L., and Doe, C.Q. (2010). Drosophila type II neuroblast lineages keep Prospero levels low to generate large clones that contribute to the adult brain central complex. Neural Dev. 5, 26.

Bello, B.C., Izergina, N., Caussinus, E., and Reichert, H. (2008). Amplification of neural stem cell proliferation by intermediate progenitor cells in Drosophila brain development. Neural Dev. 3, 5.

Boone, J.Q., and Doe, C.Q. (2008). Identification of Drosophila type II neuroblast lineages containing transit amplifying ganglion mother cells. Dev. Neurobiol. 68, 1185–1195.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. 36, 411–420.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature 566, 496–502.

Choi, H.M.T., Schwarzkopf, M., Fornace, M.E., Acharya, A., Artavanis, G., Stegmaier, J., Cunha, A., and Pierce, N.A. (2018). Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. Development 145.

Cocanougher, B.T., Wittenbach, J.D., Long, X., Kohn, A.B., Norekian, T.P., Yan, J., Colonell, J., Masson, J.-B., Truman, J.W., Cardona, A., et al. (2019). Comparative single-cell transcriptomics of complete insect nervous systems. BioRxiv.

Deitcher, D.L., Ueda, A., Stewart, B.A., Burgess, R.W., Kidokoro, Y., and Schwarz, T.L. (1998). Distinct requirements for evoked and spontaneous release of neurotransmitter are revealed by mutations in the Drosophila gene neuronal-synaptobrevin. J. Neurosci. 18, 2028–2039.

Flybase curators (2019). FlyBase Reference Report: FlyBase, 2019-, Computation of D. melanogaster genes relevant to disease based on their orthology to human “disease genes”.

Gold, K.S., and Brand, A.H. (2014). Optix defines a neuroepithelial compartment in the optic lobe of the Drosophila brain. Neural Dev. 9, 18.

Greig, L.C., Woodworth, M.B., Galazo, M.J., Padmanabhan, H., and Macklis, J.D. (2013). Molecular logic of neocortical projection neuron specification, development and diversity. Nat. Rev. Neurosci. 14, 755–769.

Habib, N., Basu, A., Avraham-Davidi, I., Burks, T., Choudhury, S.R., Aguet, F., Gelfand, E., Ardlie, K., Weitz, D.A., Rozenblatt-Rosen, O., et al. (2017). DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq. BioRxiv.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. Cell 172, 1091-1107.e17.

Homem, C.C.F., and Knoblich, J.A. (2012). Drosophila neuroblasts: a model for stem cell biology. Development 139, 4297–4310.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37, 1–13.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57.

Kalhor, R., Kalhor, K., Mejia, L., Leeper, K., Graveline, A., Mali, P., and Church, G.M. (2018). Developmental barcoding of whole mouse via homing CRISPR. Science 361.

Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. Nat. Rev. Genet. 20, 273–282.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods 16, 1289–1296.

Kose, H., Rose, D., Zhu, X., and Chiba, A. (1997). Homophilic synaptic target recognition mediated by immunoglobulin-like cell adhesion molecule Fasciclin III. Development 124, 4143–4152.

Kucherenko, M.M., Ilangovan, V., Herzig, B., Shcherbata, H.R., and Bringmann, H. (2016). TfAP-2 is required for night sleep in Drosophila. BMC Neurosci. 17, 72.

Lai, E.C., Bodner, R., and Posakony, J.W. (2000). The enhancer of split complex of Drosophila includes four Notch-regulated members of the bearded gene family. Development 127, 3441–3455.

Landskron, L., Steinmann, V., Bonnay, F., Burkard, T.R., Steinmann, J., Reichardt, I., Harzer, H., Laurenson, A.-S., Reichert, H., and Knoblich, J.A. (2018). The

asymmetrically segregating lncRNA cherub is required for transforming stem cells into malignant cells. Elife 7.

Lane, M.E., Sauer, K., Wallace, K., Jan, Y.N., Lehner, C.F., and Vaessin, H. (1996). Dacapo, a cyclin-dependent kinase inhibitor, stops cell proliferation during Drosophila development. Cell 87, 1225–1235.

Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., et al. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell 162, 184–197.

Lin, Y., Ghazanfar, S., Wang, K.Y.X., Gagnon-Bartsch, J.A., Lo, K.K., Su, X., Han, Z.-G., Ormerod, J.T., Speed, T.P., Yang, P., et al. (2019). scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. Proc Natl Acad Sci USA 116, 9775–9784.

Liu, K.S.Y., Siebert, M., Mertel, S., Knoche, E., Wegener, S., Wichmann, C., Matkovic, T., Muhammad, K., Depner, H., Mettke, C., et al. (2011). RIM-binding protein, a central part of the active zone, is essential for neurotransmitter release. Science 334, 1565–1569.

Liu, L.-Y., Long, X., Yang, C.-P., Miyares, R.L., Sugino, K., Singer, R.H., and Lee, T. (2019). Mamo decodes hierarchical temporal gradients into terminal neuronal fate. Elife 8.

Li, X., Chen, R., and Zhu, S. (2017). bHLH-O proteins balance the self-renewal and differentiation of Drosophila neural stem cells by regulating Earmuff expression. Dev. Biol. 431, 239–251.

Li, Y., Walker, L.A., Zhao, Y., Edwards, E.M., Michki, N.S., Cheng, H.P.J., Ghazzi, M., Chen, T.Y., Chen, M., Roossien, D.H., et al. (2020). Bitbow: a digital format of Brainbow enables highly efficient neuronal lineage tracing and morphology reconstruction in single brains. BioRxiv.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214.

Maier, D., Marte, B.M., Schäfer, W., Yu, Y., and Preiss, A. (1993). Drosophila evolution challenges postulated redundancy in the E(spl) gene complex. Proc Natl Acad Sci USA 90, 5464–5468.

Mariano, V., Achsel, T., Bagni, C., and Kanellopoulos, A.K. (2020). Modelling learning and memory in Drosophila to understand Intellectual Disabilities. Neuroscience.

Monastirioti, M., Giagtzoglou, N., Koumbanakis, K.A., Zacharioudaki, E., Deligiannaki,

M., Wech, I., Almeida, M., Preiss, A., Bray, S., and Delidakis, C. (2010). Drosophila Hey is a target of Notch in asymmetric divisions during embryonic and larval neurogenesis. Development 137, 191–201.

Monge, I., Krishnamurthy, R., Sims, D., Hirth, F., Spengler, M., Kammermeier, L., Reichert, H., and Mitchell, P.J. (2001). Drosophila transcription factor AP-2 in proboscis, leg and brain central complex development. Development 128, 1239–1252.

Müller, M., Genç, Ö., and Davis, G.W. (2015). RIM-binding protein links synaptic homeostasis to the stabilization and replenishment of high release probability vesicles. Neuron 85, 1056–1069.

de Nooij, J.C., Letendre, M.A., and Hariharan, I.K. (1996). A cyclin-dependent kinase inhibitor, Dacapo, is necessary for timely exit from the cell cycle during Drosophila embryogenesis. Cell 87, 1237–1247.

Ntranos, V., Yi, L., Melsted, P., and Pachter, L. (2018). Identification of transcriptional signatures for cell types from single-cell RNA-Seq. BioRxiv.

Ogawa, L.M., and Vallender, E.J. (2014). Evolutionary conservation in genes underlying human psychiatric disorders. Front. Hum. Neurosci. 8, 283.

Pei, W., Feyerabend, T.B., Rössler, J., Wang, X., Postrach, D., Busch, K., Rode, I., Klapproth, K., Dietlein, N., Quedenau, C., et al. (2017). Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. Nature 548, 456–460.

Pflugfelder, G.O., Schwarz, H., Roth, H., Poeck, B., Sigl, A., Kerscher, S., Jonschker, B., Pak, W.L., and Heisenberg, M. (1990). Genetic and molecular characterization of the optomotor-blind gene locus in Drosophila melanogaster. Genetics 126, 91–104.

Ponti, G., Obernier, K., Guinto, C., Jose, L., Bonfanti, L., and Alvarez-Buylla, A. (2013). Cell cycle and lineage progression of neural progenitors in the ventricular-subventricular zones of adult mice. Proc Natl Acad Sci USA 110, E1045-54.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods 14, 979–982.

Raj, B., Wagner, D.E., McKenna, A., Pandey, S., Klein, A.M., Shendure, J., Gagnon, J.A., and Schier, A.F. (2017). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain by scGESTALT. BioRxiv.

Ren, J., Isakova, A., Friedmann, D., Zeng, J., Grutzner, S.M., Pun, A., Zhao, G.Q., Kolluru, S.S., Wang, R., Lin, R., et al. (2019). Single-cell transcriptomes and whole-brain projections of serotonin neurons in the mouse dorsal and median raphe nuclei. Elife 8.

Ren, Q., Yang, C.-P., Liu, Z., Sugino, K., Mok, K., He, Y., Ito, M., Nern, A., Otsuna, H., and Lee, T. (2017). Stem Cell-Intrinsic, Seven-up-Triggered Temporal Factor Gradients Diversify Intermediate Neural Progenitors. Curr. Biol. 27, 1303–1313.

Saunders, A., Macosko, E., Wysoker, A., Goldman, M., Krienen, F., Bien, E., Baum, M., Wang, S., Goeva, A., Nemesh, J., et al. (2018). A Single-Cell Atlas of Cell Types, States, and Other Transcriptional Patterns from Nine Regions of the Adult Mouse Brain. BioRxiv.

Schinaman, J.M., Giesey, R.L., Mizutani, C.M., Lukacsovich, T., and Sousa-Neves, R. (2014). The KRÜPPEL-like transcription factor DATILÓGRAFO is required in specific cholinergic neurons for sexual receptivity in Drosophila females. PLoS Biol. 12, e1001964.

Servén, D., Brummitt, C., and Abedi, H. (2018). pyGAM: Generalized Additive Models in Python (Zenodo).

Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'er, D. (2019). Characterization of cell fate probabilities in single-cell data with Palantir. Nat. Biotechnol. 37, 451–460.

Snow, P.M., Bieber, A.J., and Goodman, C.S. (1989). Fasciclin III: a novel homophilic adhesion molecule in Drosophila. Cell 59, 313–323.

Soldatov, R., Kaucka, M., Kastriti, M.E., Petersen, J., Chontorotzea, T., Englmaier, L., Akkuratova, N., Yang, Y., Häring, M., Dyachuk, V., et al. (2019). Spatiotemporal structure of cell fate decisions in murine neural crest. Science 364.

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., and Junker, J.P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. Nat. Biotechnol. 36, 469–473.

Syed, M.H., Mark, B., and Doe, C.Q. (2017). Steroid hormone induction of temporal gene expression in Drosophila brain neuroblasts generates neuronal and glial diversity. Elife 6.

Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. 9, 5233.

Tran, T.N., and Bader, G. (2019). Tempora: cell trajectory inference using time-series single-cell RNA sequencing data. BioRxiv.

Tumbar, T., Guasch, G., Greco, V., Blanpain, C., Lowry, W.E., Rendl, M., and Fuchs, E. (2004). Defining the epithelial stem cell niche in skin. Science 303, 359–363.

Turek, M., Lewandrowski, I., and Bringmann, H. (2013). An AP2 transcription factor is required for a sleep-active neuron to induce sleep-like quiescence in C. elegans. Curr.

Biol. 23, 2215–2223.

Veling, M.W., Li, Y., Veling, M.T., Litts, C., Michki, N., Liu, H., Ye, B., and Cai, D. (2019). Identification of Neuronal Lineages in the Drosophila Peripheral Nervous System with a “Digital” Multi-spectral Lineage Tracing System. Cell Rep. 29, 3303-3312.e3.

Weber, T.S., Dukes, M., Miles, D.C., Glaser, S.P., Naik, S.H., and Duffy, K.R. (2016). Site-specific recombinatorics: in situ cellular barcoding with the Cre Lox system. BMC Syst. Biol. 10, 43.

Wells, R.E., Barry, J.D., Warrington, S.J., Cuhlmann, S., Evans, P., Huber, W., Strutt, D., and Zeidler, M.P. (2013). Control of tissue morphology by Fasciclin III-mediated intercellular adhesion. Development 140, 3858–3868.

Weng, M., Golden, K.L., and Lee, C.-Y. (2010). dFezf/Earmuff maintains the restricted developmental potential of intermediate neural progenitors in Drosophila. Dev. Cell 18, 126–135.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19, 15.

Yang, L., Titlow, J., Ennis, D., Smith, C., Mitchell, J., Young, F.L., Waddell, S., Ish-Horowicz, D., and Davis, I. (2017). Single molecule fluorescence in situ hybridisation for quantitating post-transcriptional regulation in Drosophila brains. Methods 126, 166–176.

Zhong, S., Zhang, S., Fan, X., Wu, Q., Yan, L., Dong, J., Zhang, H., Li, L., Sun, L., Pan, N., et al. (2018). A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. Nature 555, 524–528.

Zhou, Q., Liu, M., Xia, X., Gong, T., Feng, J., Liu, W., Liu, Y., Zhen, B., Wang, Y., Ding, C., et al. (2017). A mouse tissue transcription factor atlas. Nat. Commun. 8, 15089.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol. Cell 65, 631-643.e4.

Chapter 3

Making Single-Cell RNA-Seq Analysis Accessible For All

What does it take to complete an scRNA-seq experiment from start to finish?

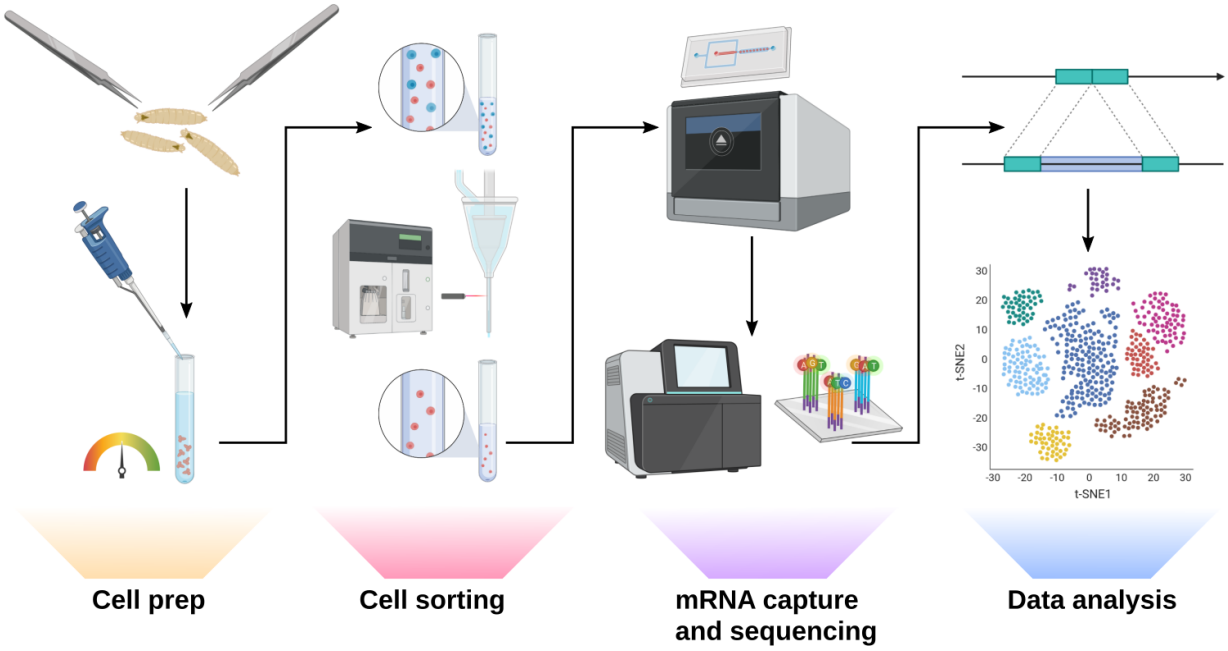

Fig. 3.1: Diagram outlining steps in a typical scRNA-seq experiment
A typical single-cell RNA-seq (scRNA-seq) experiment can be broken down into 4 main steps Cell prep involves collecting animals, dissecting out the tissue(s) of interest, and dissociating them into a single-cell suspension via a combination of chemical and mechanical tissue disruption techniques at an experimentally optimized temperature that balances enzyme activity with cell viability. Cell sorting, an optional step, ensures that only cells labelled with a genetically encoded or antibody-added dye are collected for downstream processing. Capturing the mRNA and sequencing it are at the core of these experiments, with many thesis written in recent years on methods for improving capture efficiency, experimental scale, and sequencing fidelity. Finally, data analysis (both aligning the sequencing reads to the transcriptome and processing the gene-by-

cell counts matrix) enables researchers to ultimately form and test hypotheses about the cell populations they sequenced.

Though they are by no means the most challenging type of experiment to perform in a cell biology lab, single-cell RNA-seq (scRNA-seq) experiments are complex, relying on the intersection of many experimental and analytic techniques in order to generate and test meaningful hypotheses about populations of cells at the mRNA expression level. As I show in Fig 3.1, these techniques roughly span 4 domains: sample/cell prep, cell sorting/isolation, single-cell mRNA capture and sequencing, and data analysis. Here I will describe some of the literature surrounding each of these domains, ending in the data analysis domain which I will describe in greater detail throughout this chapter.

Tissue dissociation

Unlike newer whole tissue spatial transcriptomics assays (Asp et al., 2020; Cho et al., 2021; Liu et al., 2020; Rodriques et al., 2019), traditional high-throughput scRNA-seq methods typically require that tissues be dissociated into single-cell suspensions before processing (Macosko et al., 2015; Picelli et al., 2014; Ziegenhain et al., 2017). This is fundamentally due to the fact that in order to capture and process mRNA from a single cell, a single cell must be separated from the rest of its cohort, otherwise mRNA from multiple cells would be combined and any downstream analysis of those mRNA molecules would be cofounded (i.e. we would not know which cell contributed which mRNA molecule to the pool of mRNA).

Tissue dissociation methods aim to take an intact tissue, such as the developing larval brain from *Drosophila*, and by some combination of enzymatic and mechanical action break that tissue up into its component cells. In most cases, this is done without chemical fixation, as *viable* dissociated cells are the desired end product (though some mRNA can be extracted from fixed tissues; see for example (Alles et al., 2017; Denisenko et al., 2020)). As different tissues in different organisms are made up of different component cell types, dissociation protocols vary wildly, each aiming to improve the yield and health of viable cell suspensions (Pretlow and Pretlow, 1987). In the larval and adult *Drosophila* brain others have attempted to generate viable single cell suspensions of neuroblasts and neurons for FACS and scRNA-seq studies (Croset

et al., 2017, 2018; Davie et al., 2018; Harzer et al., 2013; Li et al., 2017). I aimed to improve upon these protocols, with the specific goal of increasing the yield of neurons from the tightly-connected ventral nerve cord (VNC) to ensure fair representation of cells from both the CNS and PNS for future studies. Using a trh-Gal4 genetic driver line crossed to our UAS-hH2B::2xmNG reporter line to label all serotonergic neurons in the developing larval brain (84 in total, with cell bodies residing primarily in the VNC (Li et al., 2020b)), I systematically dissociated individual brains and counted the number of identifiable mNG+ cells in single wells of a 96-well dish post-dissociation, testing a variety of different enzymatic and mechanical dissociation conditions to obtain a protocol yielding the maximum number of viable neurons. This optimized protocol is available in greater detail in appendix A3, and intermediate optimization steps are characterized in Fig 3.2.

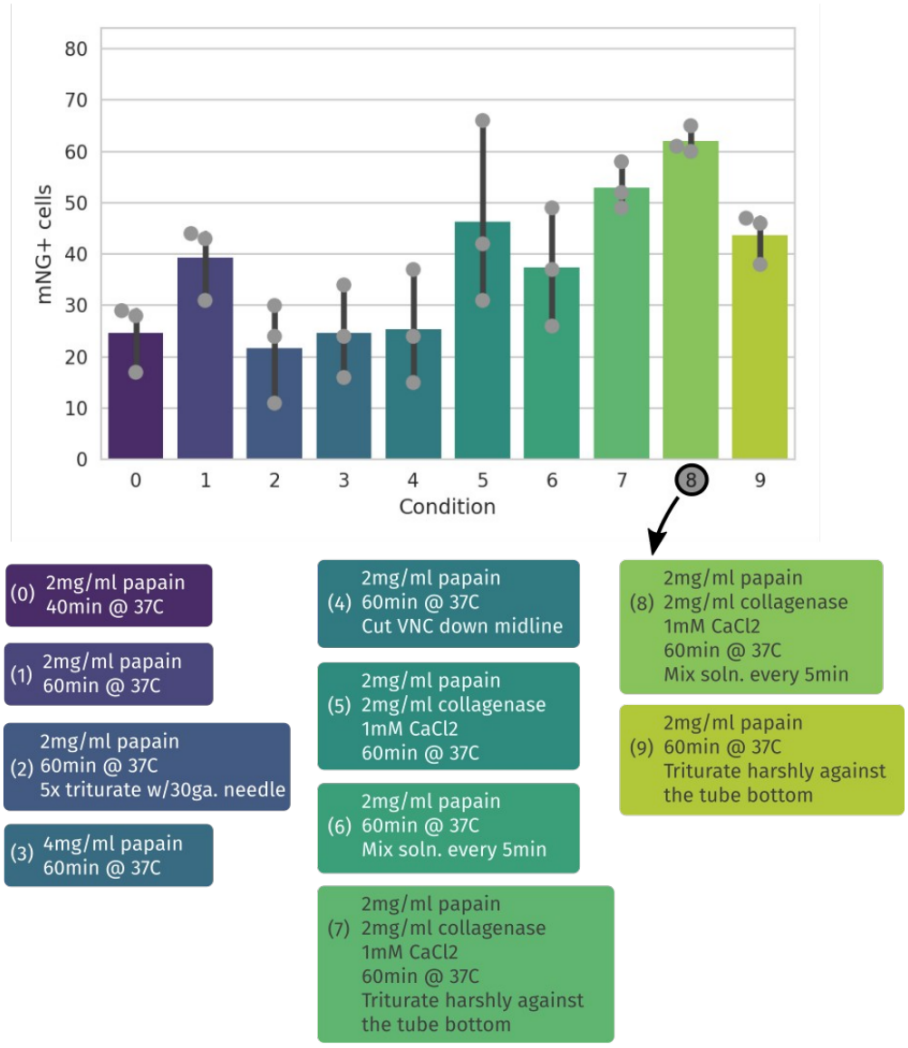

Fig 3.2: Larval brain dissociation optimizations
The number of single cells expressing mNG after dissociation from single trh-Gal4 x UAS-hH2B::2xmNG brains under a variety of dissociation conditions. Colored bars represent the mean of n=3 independent replicates (grey circles) per condition. 1mM $CaCL_2$ can be replaced with 15uM $ZnCL_2$ as both positive ions activate collagenase with similar effectiveness (and Zn^+ is generally less toxic to neurons).

In general, the combination of the proteases papain and collagenase type I, elevated temperature, and periodic mixing led to a reproducible maximum in yield of labelled trh-neurons, and so this protocol was used for all single-cell experiments outlined in this work.

Cell sorting/isolation

In order to perform targeted single-cell sequencing of specifically labelled subpopulations of cells, cell sorting/isolation techniques are required. Some methods have been developed to sort cells based on their mass/size using centrifugal forces both in test tubes (Rizzardi et al., 2016) and in microfluidic chips (Lin et al., 2017). Others utilize magnetic particles coated with antibodies that recognize cell type-specific surface antigens (Miltenyi et al., 1990).

Arguably the most flexible and popular method of cell selection is fluorescence activated cell sorting (FACS), wherein cells are individually passed in front of a laser(s) and their size (side-scatter), complexity (back-scatter), and fluorescent activity are recorded. Based on user-defined ‘gates’ around these metrics, cells with specific characteristics can be isolated apart from their cohort in rapid succession (Herzenberg et al., 2002). Recent studies (Berger et al., 2012; Harzer et al., 2013), including this work, have made use of genetic labelling schemes to incorporate bright fluorescent proteins into cells of interest in order to make them sortable on FACS machines. Surface staining of cell suspensions with dye-conjugated antibodies is also common (Herzenberg et al., 2002). Though high flow rates in FACS machines are thought to induce some cellular stress, gene expression is only minimally affected by FACS (Richardson et al., 2015), making it suitable for scRNA-seq and related studies.

mRNA capture and sequencing

After tissues have been dissociated and specific cellular subpopulations have been (optionally) isolated, the single-cell mRNA capture process can begin. Many methods have been developed to meet this need, falling into four major categories: plate-based methods, droplet-based methods, combinatorial split-and-pool methods, and hybrids of these aforementioned categories. Here I will briefly describe the general biochemical principles that *all* of these methods rely on before briefly detailing the working principles behind two of the more popular categories.

Despite some biochemical similarities, RNA cannot be recognized by the taq DNA polymerases used in PCR reactions, thereby making it impossible to amplify without first making a cDNA copy of the RNA template. In bulk RNA-seq reactions this is done by extracting RNA from many cells in a single reaction vessel and mixing this RNA with a reverse transcriptase enzyme and mRNA capture oligos - single-stranded DNA oligos the contain both a poly-dT region (complementary to the poly-A region at the 3' end of mature mRNA) and a PCR handle. Coupled with a template-switching oligo which can add a second PCR handle at the 5' end (Gilboa et al., 1979), this reaction results in the generation of a single stranded cDNA molecule that contains a sequence complementary to that of the template mRNA.

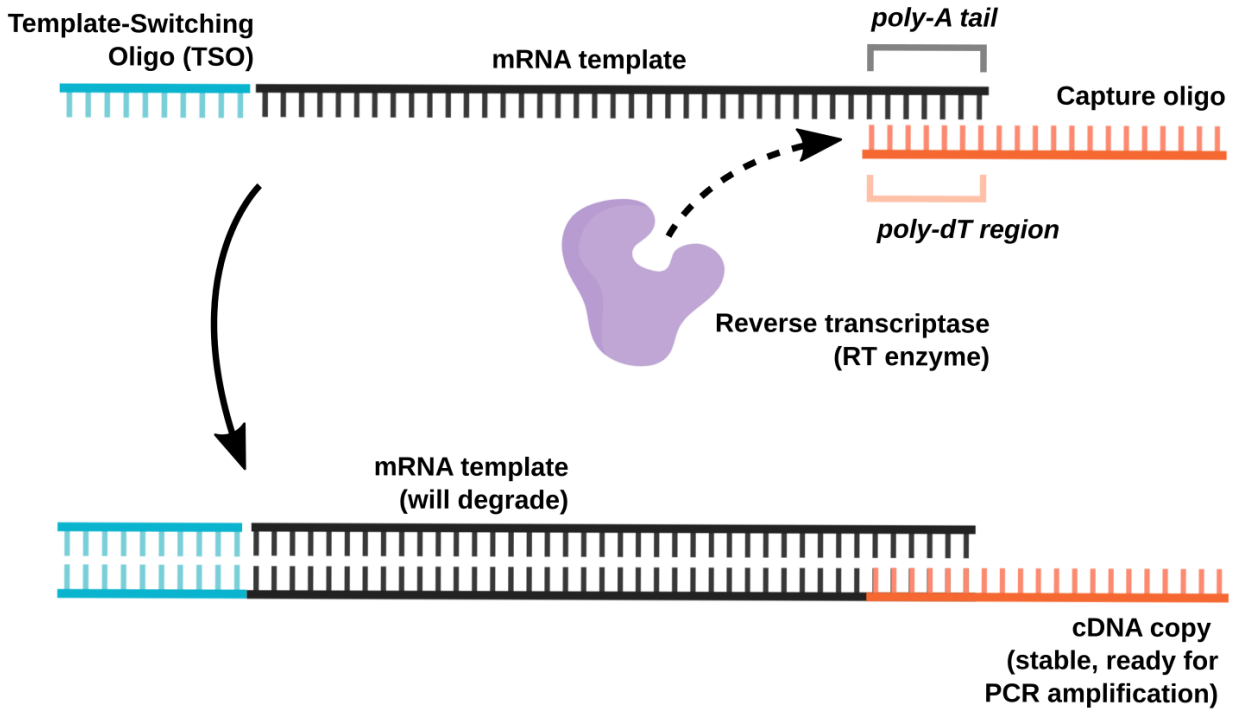

Fig. 3.3: A simplified view of a reverse-transcription reaction
The first step in most RNA-seq experiments after RNA extraction, a capture oligo that is partially complementary to the poly-A tail of the target mRNA is mixed with a reverse transcriptase (RT enzyme) and template switching oligo (TSO) in order to generate a stable cDNA copy of the template mRNA that is ready for PCR amplification.

What differentiates single-cell mRNA capture chemistries is the addition of a unique, PCR-amplifiable 'cell barcode' (and optionally a Unique Molecular Identifier/UMI) to

each mRNA template. These sequences are typically added to the poly-dT capture oligo as in (Macosko et al., 2015), though they can alternatively be added through direction ligation, as in (Hochgerner et al., 2017), or in later sequencing library preparation steps using unique sequencing library adapters, as in (Picelli et al., 2014). At a high level, the end result is the same - each mRNA molecule is uniquely tagged by a UMI and a cell barcode that is unique to each cell but shared by all mRNA from that cell.

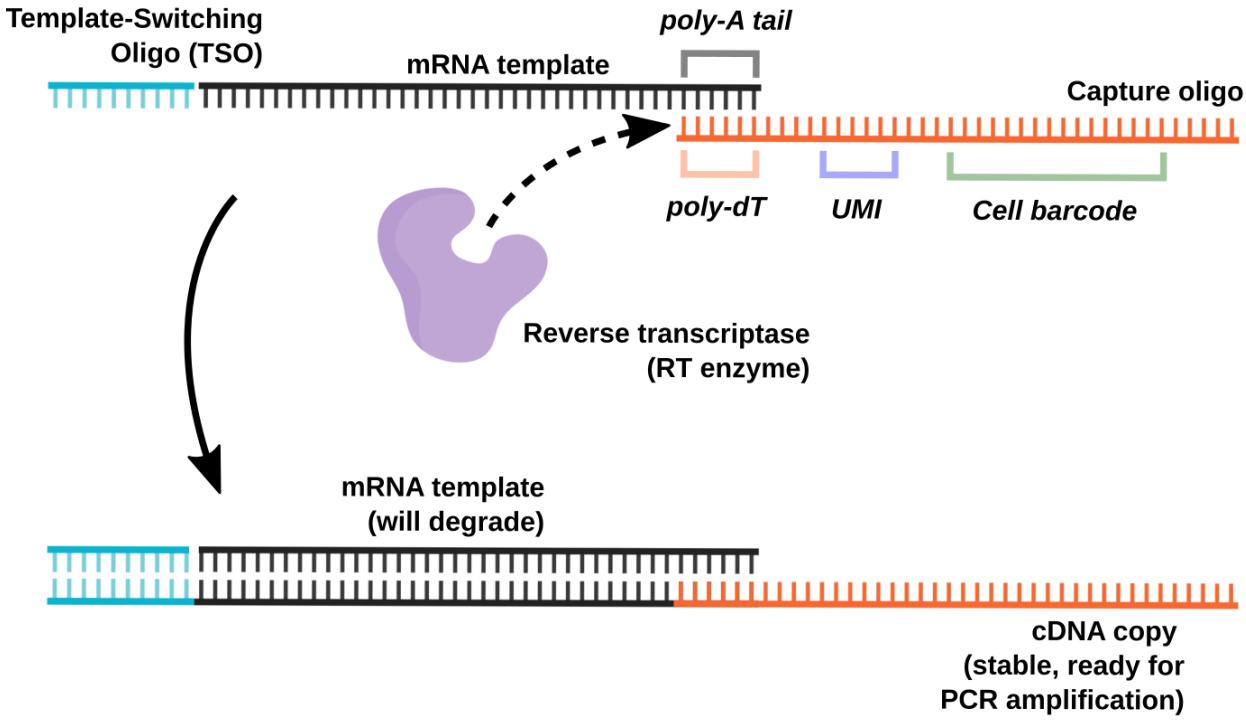

Fig. 3.4: An RT reaction with cell barcodes and UMIs
The addition of Unique Molecular Identifier (UMI, a stretch of random nucleotides unique to each capture oligo) and Cell Barcode sequences (a stretch of random nucleotides unique to each *cell*) to the poly-dT capture olgo, in combination with single-cell isolation techniques, biochemically enables assignment of each mRNA molecule to a cell of origin. UMIs enable correction for PCR amplification biases by providing a unique identifier for each molecule *before* those molecules are amplified by PCR in preparation for sequencing - important for accurately modelling gene expression levels in scRNA-seq experiments.

The addition of cell barcodes to poly-dT capture probes is not particularly useful without physically isolating individual cells from one another in separate reaction vessels, in order to react capture oligos containing one cell barcode with mRNA from only one cell.

The simplest (and typically most costly) scRNA-seq protocols involve sorting single cells into individual wells of multi-well PCR plates that can be assigned unique cell barcodes at the final sequencing library preparation step (Hagemann-Jensen et al., 2020; Picelli et al., 2014). These techniques benefit from a massive surplus of reaction volume and the ability to capture and process full-length mRNA transcripts, making them powerful for the detection of rare mRNAs and isoforms.

To massively improve cell throughput, microfluidic droplet-based scRNA-seq methods have been developed and popularized through both in-house and commercial solutions (Habib et al., 2017; Macosko et al., 2015; Ziegenhain et al., 2017). These methods co-encapsulate single cells with beads containing poly-dT capture oligos inside of water-in-oil droplet emulsions at a rate of thousands of droplets per second. This droplet emulsion can then be placed in a single reaction vessel (tube), thousands of cells now reacting independently with capture oligos from their own unique bead (as the oil prevents cross-talk between droplets). Without a doubt the most popular solution thanks in part to commercialization work by 10X Genomics, more than 2500 papers have been published using technologies of this type since 2015 (Saxonov, 10X Genomics Investor Presentation, 2021).

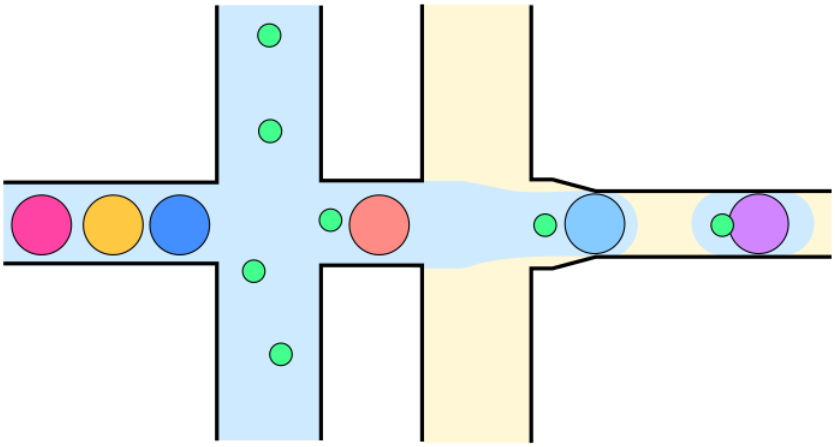

Fig. 3.5: Simplified diagram of a 'Drop-seq'-style single cell mRNA capture device
Beads (large circles) are co-flowed with cells (small circles) and pushed from left to right through the microfluidic device. Droplets are generated by introducing an inert oil from a third channel at a high flow rate and pushing the mixture through a narrow junction, essentially 'pinching off' the bead/cell mixture and forming droplets separated by oil.

Each droplet then is its own independent reaction vessel, and as cells are lysed in each droplet their mRNA will hybridize to capture oligos provided by each bead, with oil preventing mRNA from transferring between droplets.

After capturing, barcoding, and generating amplified cDNA copies of the mRNA from each cell, molecules need to be sequenced. With the advent of next-generation sequencing technologies in the mid-2000s coming about first by the development and commercialization of sequencing-by-ligation (Shendure et al., 2005) technologies and subsequently launched forward by the development and commercialization of sequencing-by-synthesis (Bentley et al., 2008) technologies (Illumina-style sequencers), generating more than 400 million high-fidelity sequencing reads from a scRNA-seq experiment is now routine. Extremely powerful for assaying gene expression at the mRNA level, the main limitation of these popular Illumina-style sequencers is their short read lengths, with most scRNA-seq libraries generating data from <100bp of each captured mRNA molecule, though reads as long as 600bp are commercially supported (though significantly more costly). Since most mRNA transcripts are larger than 2kb, short reads can limit the ability to detect alternatively spliced mRNA molecules coming from the same gene locus, making RNA-velocity (La Manno et al., 2018) and isoform-specific (for example, (Joglekar et al., 2021)) analyses challenging.

Data analysis - sequence alignment

Typical scRNA-seq datasets comprise more than 400 million sequencing reads, each of which might represent a portion of an mRNA transcript. In order to determine *which* gene each sequencing read corresponds to, the first step in most scRNA-seq data analysis pipelines is to align/map reads to the transcriptome of the organism(s) of origin. Though computationally taxing, many tools are available to accomplish this task for most organisms, with STAR (Dobin et al., 2013; Kaminow et al., 2021), CellRanger (10X Genomics), Salmon (Patro et al., 2017), and Kallisto-bustools (Bray et al., 2016; Melsted et al., 2021) being popular options. A great deal of effort has gone into the development and optimization of these tools, and it is beyond the scope of my thesis to describe them in detail. Irrespective of their algorithmic differences, each tool can

generate a so-called ‘counts matrix’, with unique cell barcodes and gene names making up the rows/columns, respectively. Gene expression values in the form of counts populate the matrix, representing the (typically integer) number of unique mRNA transcripts belonging to each specific gene from each unique cell.

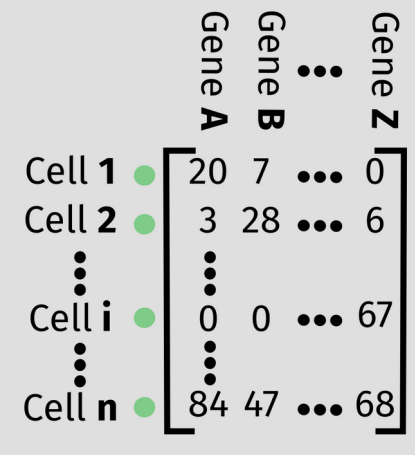

Fig. 3.6: A diagram of a typical scRNA-seq counts matrix
Unique cell IDs (cell barcodes) make up the row labels, while gene IDs make up the column labels. Integer counts populate the matrix, representing the number of sequencing reads that correspond to a specific gene in a unique cell. In scRNA-seq protocols that include unique molecular identifiers (UMIs), these gene expression counts represent truly unique mRNA molecules. Without UMIs, these gene expression counts may include PCR duplicates, over-estimating the true mRNA expression level.

Data analysis - secondary/downstream analysis

On its own, the cell-by-gene counts matrix is not particularly useful. An attempt to represent the mapping between raw sequencing reads to gene expression levels in thousands of cells at a time, these matrices are large, limited by mRNA sampling rates (so-called ‘dropouts’ leading to extremely sparse matrices (Cao et al.; Svensson, 2019)), and difficult to interpret on their own. Methods from machine learning and linear algebra have been co-opted to help researchers filter and reduce the complexity of these datasets, and here I refer to this collection of methods as secondary or downstream scRNA-seq analysis (in contrast to sequence alignment).

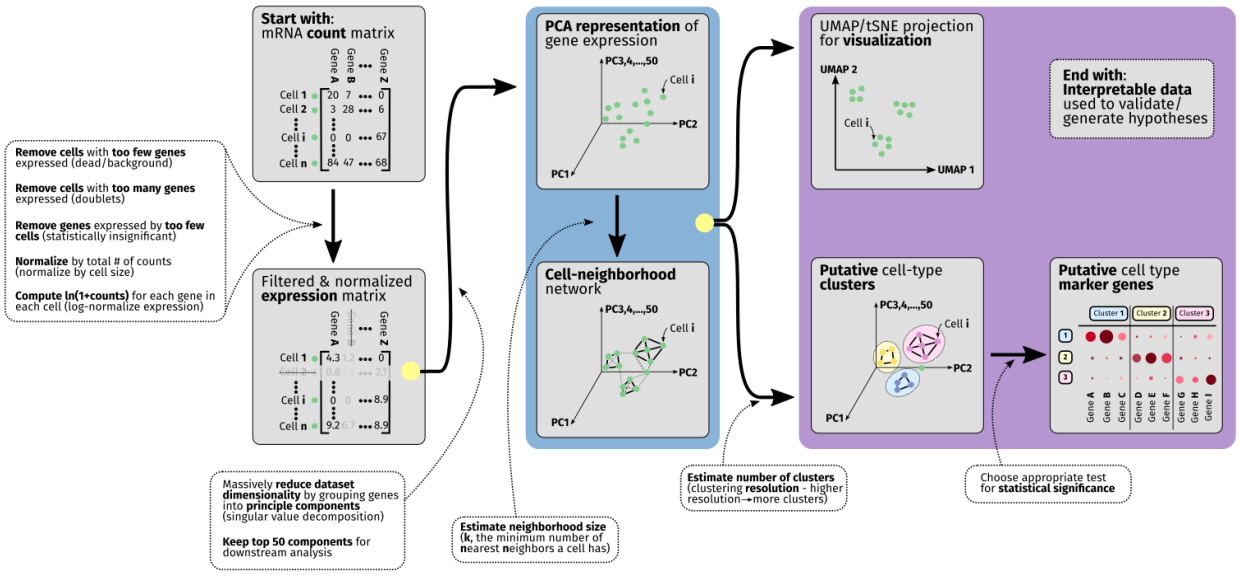

Fig. 3.7: A typical secondary/downstream scRNA-seq analysis pipeline/protocol Dashed boxes/arrows indicate methods that are applied to transform the data from a raw counts matrix into a set of interpretable plots and metrics that researchers can use to draw conclusions about how heterogeneous their cell population(s) are and what genes are expressed to differentiate these cell types from one another. With more than 10 transformation steps, each with adjustable parameters, these protocols can be challenging to work through. Solid arrows here indicate dependencies; for example, UMAP and cell type clustering analysis both require that cell neighborhood and PCA representations of the data be computed, but do not depend on one another.

Developing a Multi-informatic Cellular Visualization tool (MiCV) to make secondary scRNA-seq analysis accessible to all

Though many toolkits have been developed to perform the steps in a typical scRNA-seq secondary analysis pipeline (Dhapola et al., 2021; Li et al., 2020a; Lopez et al., 2018; McCarthy et al., 2017; Satija et al., 2015; Wolf et al., 2018), little work has been done to provide easy-to-use graphical applications based on these toolkits. This makes scRNA-seq analysis challenging for many researchers whose expertise are not in coding with advanced programming languages like R and python, but in their experimental techniques and deep knowledge of the literature context surrounding their scRNA-seq data. This lack of graphical tools does not come as a great surprise, as developing user-facing graphical applications requires a tremendous amount of programming effort

largely on for rudimentary tasks of little academic significance (for example, code to place a button on a screen at a specific location, and related code to make that button trigger an action when pressed, but only when a user is logged in, and so forth). Typically such programming tasks stay in the commercial realm, but companies starting from scratch in the area of scRNA-seq secondary analysis often ‘reinvent the wheel’ instead of building on top of the great open source academic developments outlined previously. With this in mind, I developed MiCV (Michki et al., 2021), a web-based graphical tool that enables researchers with no programming experience to leverage these toolkits for their own analyses. Building on top of scanpy (Wolf et al., 2018), a python-based scRNA-seq secondary analysis toolkit, alongside many other open source libraries, MiCV exposes researchers without any programming experience to an analysis pipeline that keeps up with trends in the rapidly developing bioinformatics field and makes it possible for a single push of a button to generate publication-ready plots and analyses. A mix of interactive and static plots and reports make iterating on cluster assignments, marker gene identification, and pseudotime analyses possible in minutes instead of days.

Here, I will outline the main steps in a typical scRNA-seq secondary analysis pipeline, and outline alongside each step some features of the MiCV web tool that allow users to interact with this pipeline as they see fit.

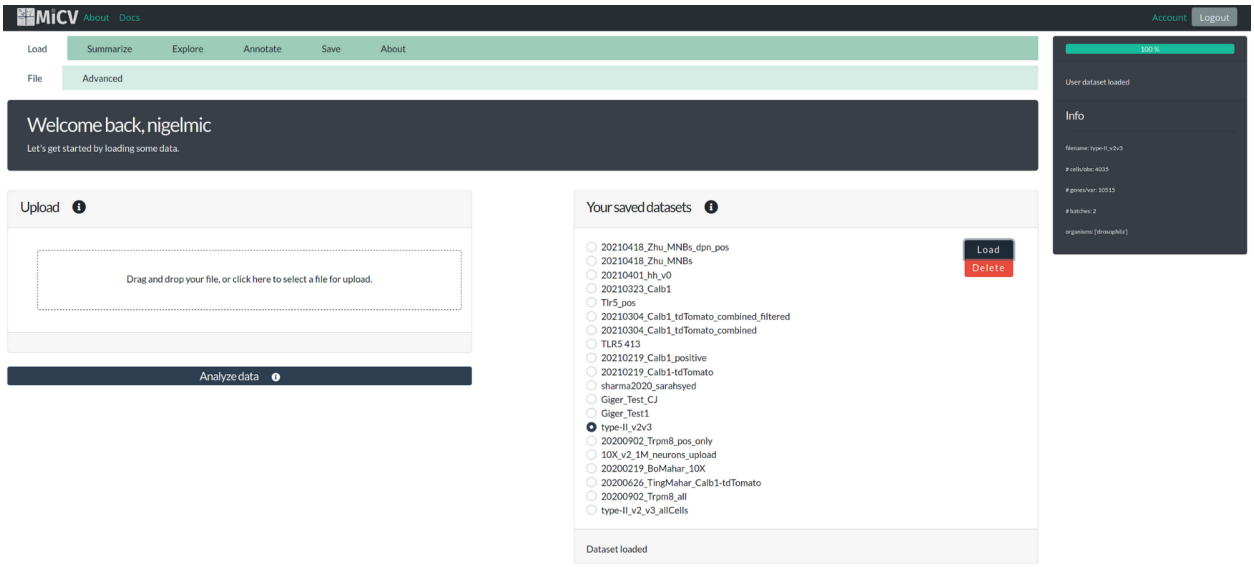

Fig. 3.8: The first page of the MiCV web tool
MiCV's post-login page provides options to upload data or load previously-analyzed datasets. A status panel on the right informs users about what operations, if any, are currently on-going in the background or have been completed previously. Users are encouraged to *not* change default pipeline parameters by only being provided with an 'Analyze data' button on this page, with more advanced controls hidden away in a separate 'Advanced' tab. This makes the tool easier to use (there is only one button to press to move through the entire secondary analysis pipeline), improves reproducibility (by having many users make use of the same default parameters), and still enables more advanced use-cases without the need for other tools.

Filtering the counts matrix

The first step in most scRNA-seq downstream analysis pipelines involves filtering the counts matrix. Here, cells that express too many or too few genes/UMI counts/reads are removed from the dataset, likely representing 'doublets' (cells that were physically co-encapsulated into the same reaction vessel, and thus have the same cell barcode) and low-quality cells and/or cell barcodes with only background mRNA, respectively. So-called 'knee-plots' are often used to determine lower bounds for this cell filtering step; when cell barcodes are ranked based on the number of UMI counts they were assigned, a sudden dropoff is usually identifiable. Cell barcodes with fewer UMIs than this inflection point (knee) are typically believed to represent low quality cells and

background RNA, though some debate in the field remains as to how strictly this cutoff should be followed (Lun et al., 2019), as some cell barcodes with very few UMIs may represent true cell types that simply express far less mRNA than others in the dataset.

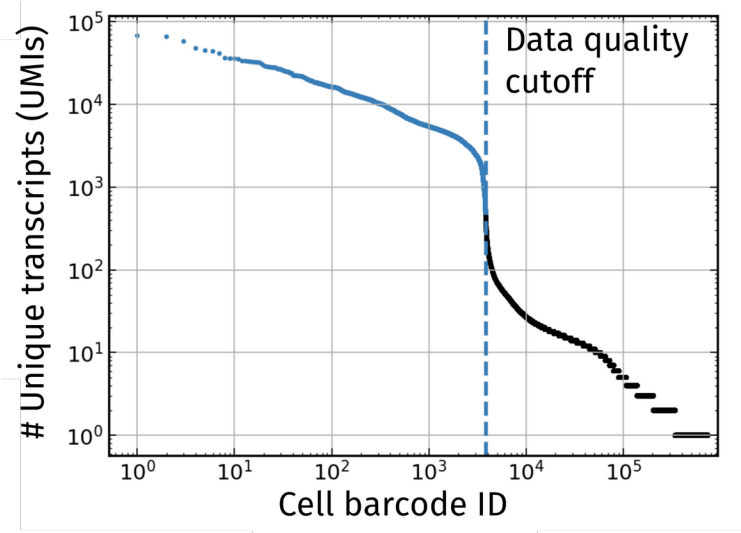

Fig. 3.9: A 'knee-plot', showing the number of unique mRNA transcripts (UMIs) associated with each cell barcode
When plotted on log-log scales, a sudden inflection point can typically be identified (vertical dashed line); cells to the right of this line are considered low-quality or representative of background mRNA, and are thus removed from the dataset. Cells with far more UMIs than the median may also be filtered, though this cutoff point is much more arbitrary and relies heavily on knowledge about the cell type distribution the data represents (for example, actively proliferating cells may express far more mRNA than quiescent cells, and so a wide range of UMI counts should be expected). Data in this plot comes from the 10Xv2 type II scRNA-seq experiment described in chapter 2, and was generated using UMI-tools (Smith et al., 2017).

Genes are also filtered, removed here if they are expressed in too few cells for statistical comparisons to be meaningful. In practice, many tutorials recommend removing genes expressed in fewer than 2-5 cells (Satija et al., 2015; Wolf et al., 2018), though cutoffs expressed as a fraction of the cell count (for example, 0.01% of all cells) may be more appropriate for datasets of varying size. Using MiCV, these filtering parameters can be modified from their preset defaults using slider bars that indicate relevant ranges of parameter values for users to explore.

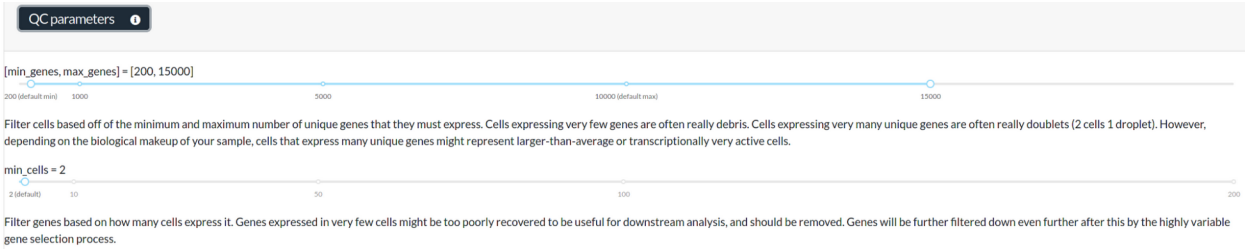

Figure 3.10: MiCV interface for modifying cell/gene filtering parameters
Defaults are marked along the slider bars, and currently selected values are indicated in the text above each slider to remove ambiguity and improve reproducibility. Similar interfaces are provided for selecting the number of highly variable genes, the number of neighbors in the cell-cell neighborhood network, and the clustering resolution.

Converting the counts matrix into a gene expression matrix (normalization)

In order to model the noise distribution of this filtered counts matrix and stabilize its variance, a variety of gene expression normalization strategies are used. Typically, estimating size factors on a per-cell basis, using these size factors to normalize each cell to have the same number of total counts (conversion to counts-per-million/CPM), and a log-transforming the data (ln(1+CPM)) accomplishes this goal (Love et al., 2014) and has been used before to compare bulk RNA-seq experiments. Other methods have also been proposed, though their increased complexity has limited their adoption (Grün et al., 2014; Hafemeister and Satija, 2019; Lopez et al., 2018; Svensson et al., 2020), as have debates in the field about which noise model(s) represent the true noise observed in scRNA-seq data originating from droplet-based experiments (Svensson, 2019). In MiCV, I opted to use size-factor based normalization with log-transformation, and no user-modifiable parameters are exposed to the user in order to increase reproducibility and simplicity.

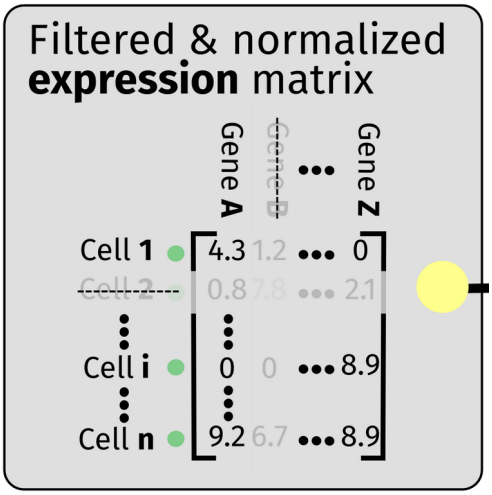

Fig. 3.11: A diagram showing the mRNA expression matrix
This is the counts matrix after filtering cells, filtering genes, normalizing expression using cell size-factors, and log-transforming the data. I call this matrix the 'expression matrix', in contrast to the counts matrix from before, as the values that populate this matrix can be fairly compared across cells and can be brought forward for further data dimensionality reduction and processing.

Reducing data dimensionality and identifying putative cell-type clusters

This filtered gene expression matrix is still large and sparse, making the identification of cell types/clusters challenging both by eye and with the help of clustering tools (Traag et al., 2019). To overcome this challenge, dimensionality reduction techniques are employed to drastically reduce the dataset dimensionality. Principal component analysis (PCA) is the most commonly used technique in this class, being closely related to familiar eigenvector/eigenvalue decomposition and/or singular value decomposition operations used in a range of linear algebra techniques (Pearson, 1901). Other tools such as scVI (Lopez et al., 2018; Svensson et al., 2020) reduce dimensionality by training neural network models to identify the 'latent space/manifold' that the gene expression data exists in/on, and data diffusion models such as those proposed by (van Dijk et al., 2018) similarly reduce data down into 'diffusion components' that represent the latent gene expression space. Though potentially more accurate, their increased complexity and computational expense make them less popular, especially for basic or first-passes at scRNA-seq secondary analysis.

Intuitively, we can think of PCA as identifying the primary weighted sets of genes that describe maximal variance in the dataset. For instance, if the genes *D*, *grh*, and *ey* in an scRNA-seq dataset from the fruit fly were responsible for contributing the most to differences (variances) in cell transcriptomes, then they might be represented in the first principal component (PC) of the PCA-reduced dataset (with some weights attached to each gene to represent their relative contributions to that variance). Genes that contribute very little to cell variance (a classic example is *Actin*, found in nearly all cells) will be represented in each PC with extremely small weights, denoting their relatively small contribution to cell-by-cell differences in gene expression (despite being a very highly expressed gene overall). While a variable number of principal components can be used, in MiCV we keep the *m*=50 top PCs, which represent the majority fraction of variance in most scRNA-seq datasets. This parameter is informed in large part by tutorials provided by the authors of scanpy (Wolf et al., 2018) and may be made user-adjustable in future versions of MiCV.

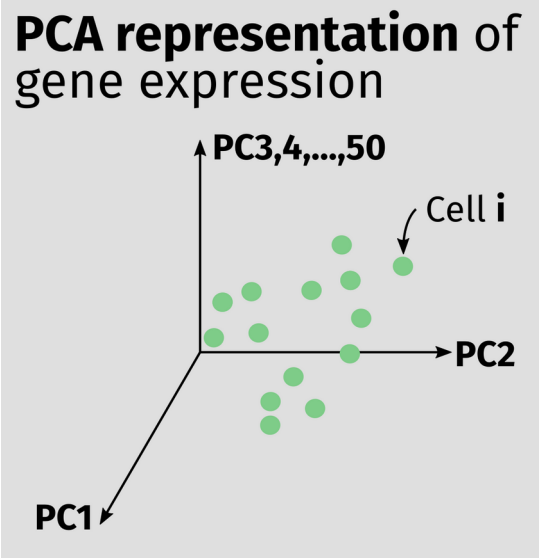

Fig. 3.12: A diagram of a PCA-reduced scRNA-seq dataset
Cells (represented as circles) are placed in an *m*-dimensional PCA space (here, m=50, though only 3 axes are represented graphically), where each cell's position along the *m* axes represents their gene expression state. Before PCA, each cell's gene expression state was represented by an *n*-dimensional space, where each dimension represents

the expression of one gene. Since for most organisms and scRNA-seq experiments $n>10,000>>m$, reducing the dimensionality of the gene expression data massively reduces noise and improves computational efficiency at the cost of some information loss during neighborhood identification and low-dimensional data embedding.

In order both the cluster cells based on their gene expression state and represent our now 50-dimensional data in 2D (or 3D) projections, a cell-cell neighborhood graph needs to be generated. Graph generation algorithms attempt to characterize the abstract structure of single-cell gene expression data by drawing edges between cells that have similar gene expression characteristics. Many algorithms to build this graph exist (Goldberger et al., 2004), though the algorithm proposed and implemented alongside the UMAP algorithm is commonly used (McInnes et al., 2018). While a bit arcane, this network description of an scRNA-seq dataset makes identifying putative cell subtype clusters possible. The main parameter that can be varied in this algorithm is k, the expected number of nearest neighbors for each cell. Intuitively, we can think of this parameter as estimating the size of the smallest cell type cluster (based on gene expression) in any given scRNA-seq dataset. Extremely large datasets may benefit from the selection of a higher value for k, due to their relatively large number of cells; however in practice a small value such as k=20 will result in similar cell-type clustering and UMAP projection results as a larger value, and is the default in MiCV and other pipelines (Wolf et al., 2018), though in MiCV this is a user-editable parameter.

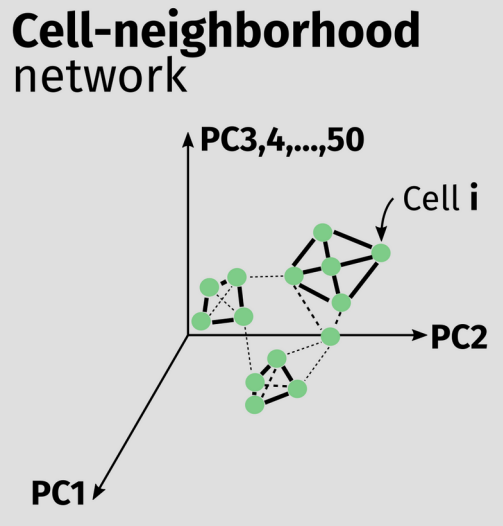

Fig. 3.13: A diagram of a cell-cell network graph
Cells are represented by circles, and the network is represented by edges of varying weights between cells projected in PCA-space. Light, dashed lines represent particularly weak connections between cells, indicating that two cells are likely not 'similar' enough to be clustered/projected together.

In order to easily visualize the complex, high-dimensional gene expression data encapsulated by the cell by gene expression matrix, low (2 or 3) dimensional 'embeddings' of the data need to be generated. Intuitively, what we aim for is a 2D plot where cells that are 'similar' in their gene expression levels are nearby to one another in 2D space, and cells that are very 'dissimilar' are farther away. Though made up of more than 10,000 dimensions (1 for each gene), methods such as tSNE (van der Maaten and Hinton, 2008) and UMAP (McInnes et al., 2018) are popular among single-cell researchers for reducing their high-dimensional gene expression data into interpretable low-dimensional plots, with UMAP gaining widespread adoption due to its increased stability. In MiCV, only UMAPs are supported, and no parameters are changed from the defaults provided by the scanpy and UMAP-learn libraries (McInnes et al., 2018; Wolf et al., 2018).

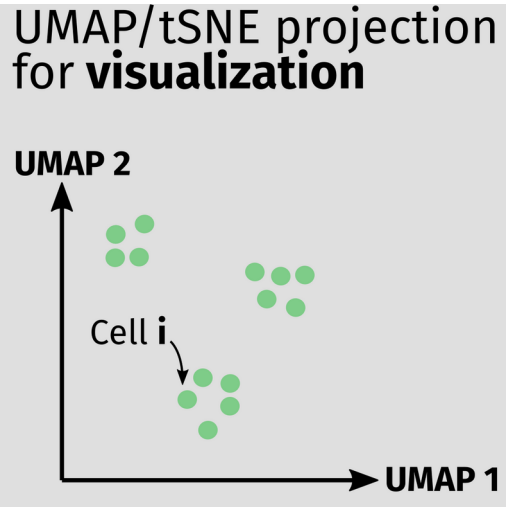

Fig. 3.14: A diagram of a typical 2D UMAP projection
Cells are more compactly clustered in 2D-space based on their local neighborhood network connections, which UMAP (and tSNE) algorithms can take into consideration when defining a high-to-low dimensional embedding of scRNA-seq data.

Though low-dimensional embeddings can impart some visual 'structure' to complex, high-dimensional scRNA-seq datasets, robustly identifying cell-type clusters requires the use of higher-dimensional representations of the data. The leiden/louvain clustering algorithms (Traag et al., 2019), used by default in scanpy and MiCV, walks the previously calculated cell-cell neighborhood network and identifies groups (clusters) of cells that are tightly connected in this graph. These well-connected groups are separated out as cell-type clusters, enabling us to do pseudo-bulk comparisons between putative cell types. We commonly label each cell in UMAP projections with colors that correspond to each cell-type in order to visualize where specific cell-types project to in a low-dimensional embedding and to build an intuition about how similar certain cell-types may be. The leiden algorithm typically can accept a user-provided *resolution* parameter that can bias the number of cell types that are identified. Higher values (>1) for this parameter will generate more, finer-grained clusters than the default, and smaller (<1) values will generate fewer, coarse-grained clusters. Both can be informative about cell type diversity, as low resolution clustering can be used to identify primary cell types (for example, 'neurons' vs. 'glia') and high resolution clustering can further break down these large groups into secondary cell types (for example,

‘excitatory neurons’ vs. ‘inhibitory neurons’). In MiCV, this is a user-definable parameter that can be used to rapidly iterate on the clustering resolution.

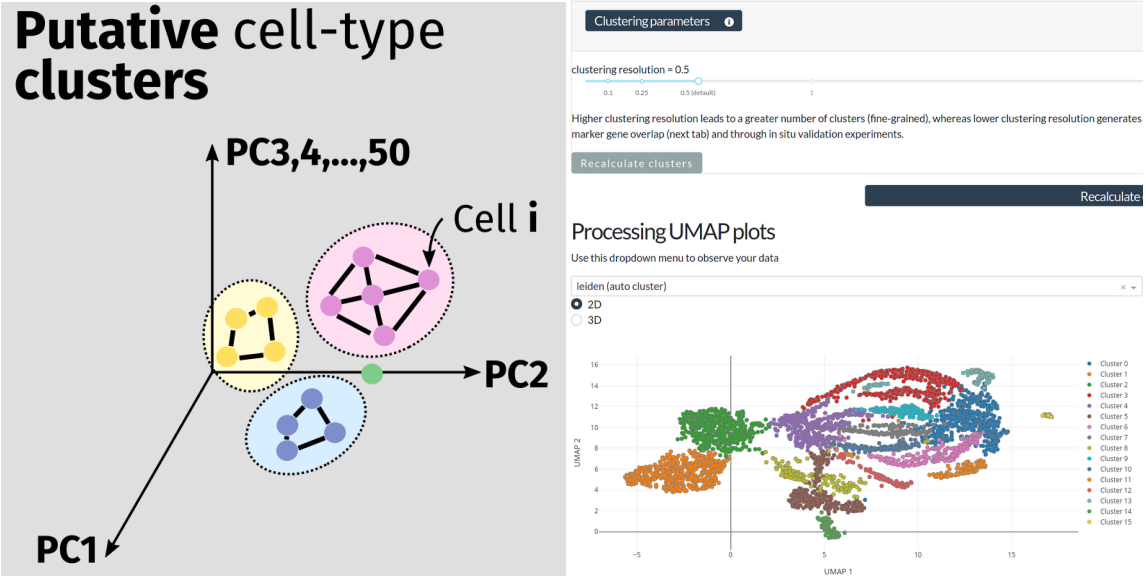

Fig. 3.15: Cell type clustering in MiCV
Left: A post-clustering diagram of the cell-cell network in PCA space. The network has been pruned of low-weight connections by the clustering algorithm, and separated networks of cells are labelled as different cell types (colors). Many graph clustering algorithms are used in network analysis, though the leiden algorithm (Traag et al., 2019) is popular for scRNA-seq data. Right: A section of the MiCV interface devoted to changing the clustering resolution parameter. Without needing to perform any PCA, neighborhood, or UMAP calculations, users can rapidly change the clustering resolution and view cluster assignments on a 2D UMAP projection by using a slider bar interface.

Identifying what makes each cell-type unique (marker gene analysis)

One of the first questions we can ask when presented with clustered scRNA-seq data is “What makes these clusters unique?”. Differential expression (DE) analysis can be used to answer this question by comparing the gene expression profiles of cells falling into each cell type cluster and identifying differentially expressed genes (DEGs) amongst the cell types clusters. Traditionally, statistical tests such as Student’s t-test or Wilcoxon rank-sum tests are used to perform these comparisons and identify genes that are up-regulated in one putative cell type cluster over the others, considering the distribution of gene expression values across cells belonging to each group (Love et al., 2014). Other

tests, such as those that overestimate the variance of each group before performing the Student's t-test (Wolf et al., 2018), perform logistic regressions (Ntranos et al., 2018), or use complex (but computationally efficient) cell binning strategies (Vuong et al., 2020), are also becoming more common. Regardless of method, these marker genes are traditionally viewed in plaintext tables and matrix or dot-plots that can compactly represent both the mean expression of each marker gene in the group and the number of cells within the group that express the gene at all. In this way, researchers can quickly identify genes of interest for specific putative cell types and validate them using *in situ* staining techniques and functional studies to test whether or not specific cell types *require* the expression of specific genes in order to function/develop correctly.

Marker-gene based cluster analysis is often where a great deal of time is spent in the scRNA-seq secondary analysis pipeline, not due to computational complexity but instead due to the need to incorporate literature studies in order to add meaningful context to marker gene lists. Being presented with a list of 10 gene names, each representing a gene that is up-regulated in a specific cell type vs. all others, can be challenging to interpret if a researcher is unfamiliar with those 10 specific genes. Considering most genomes have well over 10,000 protein-coding genes, the odds that a researcher will be familiar with any given gene are low. As such, MiCV devotes an entirely separate set of tabs/webpages to the marker gene discovery and exploration process, enabling researchers to rapidly perform marker gene analysis between subsets of putative cell types, download tables and plots representing the results therein, plot the expression of individual or groups of genes in UMAP-space, and view gene annotations/descriptions from NCBI entrez databases (Maglott et al., 2005) without needing to leave the tool or write code themselves.

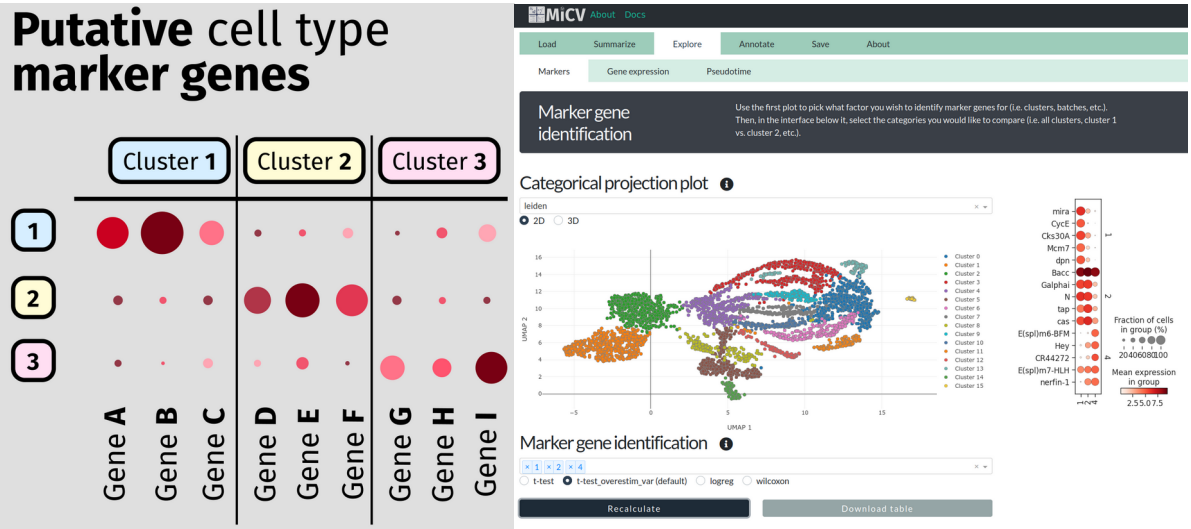

Fig. 3.16: DEG analysis in MiCV

Left: After putative cell type clustering, identifying lists of genes that are up/down-regulated in each group relative to all others is generally desirable. Dotplots (drawn here) use color intensity to indicate the mean expression level of a gene in each group and use circle (dot) size to indicate the fraction of cells in each group that express the gene at a non-zero level. Right: Since marker gene analysis is so critical to scRNA-seq analysis, MiCV devotes an entire tab/webpage to performing and visualizing the results of cell type cluster DEG analysis.

Fig. 3.17: Gene expression exploration in MiCV

In order to rapidly visualize and add literature context to the marker gene discovery process, MiCV provides a tab/webpage devoted to plotting gene expression in UMAP-space (top-left) and as 1-D value distributions/violin plots (bottom). Data from (and linkouts to) NCBI entrez databases (top-right) enable researchers to quickly come up to speed on the functional characteristics of their genes of interest, previous names/aliases, and orthologs found in other organisms. Multicolor gene expression plots, like the one shown here in the top-left, are not yet common in scRNA-seq secondary analysis toolkits, but have been introduced both in MiCV and other interactive tools such as Partek Flow (Partek Inc., 2020) as a way to compare gene expression across cell types in a similar manner to how multiplex *in situ* staining experiments are viewed. In MiCV, gene expression values are normalized to have a range of [0,1] on a per-gene basis, and so absolute expression levels of each gene are lost in these multicolor plots.

Incorporating pseudotime analysis to model developmental trajectories

Though less common than cell type clustering analysis, pseudotime techniques can be particularly useful for describing semi-continuous differentiation/degradation processes

that may be represented in scRNA-seq datasets. Also called trajectory inference methods (Saelens et al., 2019), we can intuitively think of these methods as a way to make a line of cells that is ordered based on the set of least changes from cell *i* to cell *i+1*. Given a starting cell (selected based on the expression of genes at the top of a developmental trajectory - for example, high expression of *CycE* in scRNA-seq data from the type II NB system in *Drosophila* (Michki et al., 2021) which indicates progenitor status), pseudotime trajectory inference methods search for the cell that is most similar in gene expression state to the starting cell and assign it a a pseudotime slightly higher than that of the starting cell. Then, the process repeats, searching for the cell that is most similar to the most recently placed cell in the trajectory, until all cells are aligned along the trajectory. *Many* different methods exist for fitting these trajectories (Lange et al., 2020; Qiu et al., 2017; Saelens et al., 2019; Setty et al., 2019; Street et al., 2018; Tran and Bader, 2019), each with their own algorithmic nuances and computational speedups. No one-size-fits-all method has made an obvious appearance in the field, though Monocle (Qiu et al., 2017) and Palantir (Setty et al., 2019) are popular solutions in R and python, in large part due to their comprehensive tutorials and relatively easy-to-use toolkits.

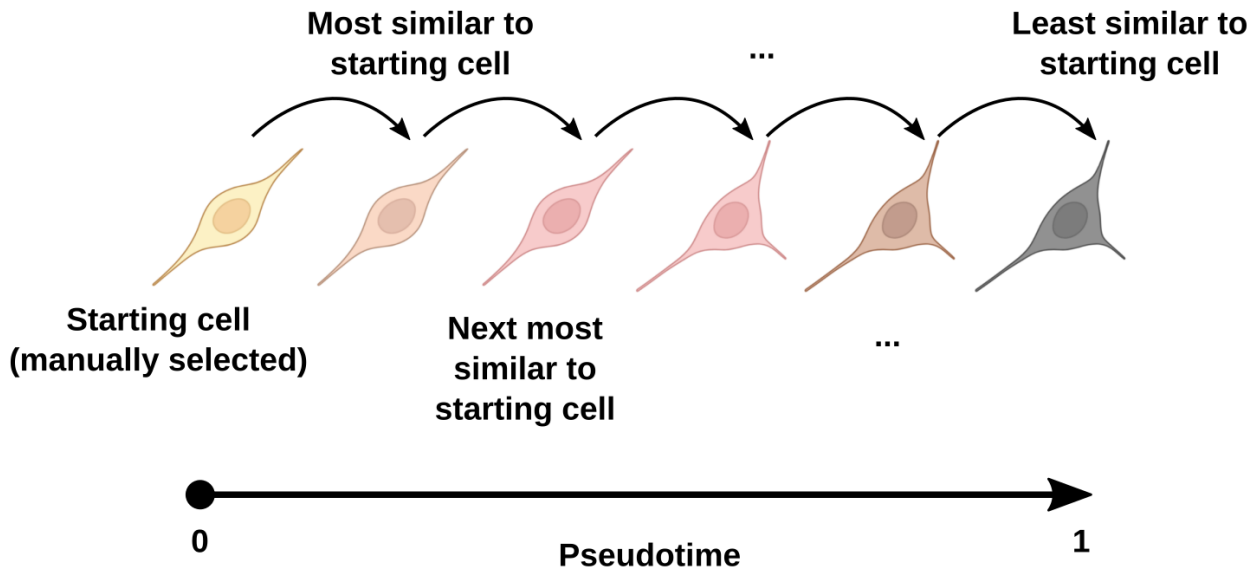

Fig. 3.18: A diagram outlining a simplified view of pseudotime trajectory inference methods
After selecting a starting cell, pseudotime methods attempt to order cells based on the set of fewest changes from cell to cell, ending with the cell(s) that are most dissimilar to the starting cell. Some methods also identify branching points in these trajectories; points where serious breaks in similarity occur, usually indicating the development of unique cellular lineages (not pictured here).

Regardless of the choice of algorithm, one of the main advantages of incorporating pseudotime analysis into an scRNA-seq secondary analysis pipeline is the ability to estimate gene expression trends across pseudotime - in a sense, giving us a view as to how gene expression changes as cells differentiate/degrade/change along the trajectory. These gene expression trends can be fit to cells along pseudotime using simple binning techniques or smoother, more complex generative additive model (GAM) fitting (Michki et al., 2021; Setty et al., 2019), and can additionally be generated in lineage branch specific manners (if multiple branch points have been identified in the pseudotime trajectory). In this way, researchers can view not only the discrete marker gene lists that delineate different putative cell types, but also the more continuous gene expression changes that oversee complex differentiation processes. These fitting methods are somewhat hindered by data non-uniformity (i.e. cells are not necessarily uniformly spread along pseudotime, but may ‘clump’ together in certain regions), a

recommended reliance on smoothed/imputed gene expression data (van Dijk et al., 2018) (increasing the risk of over-fitting) and severe dependence on GAM fitting parameters (such as the number and degree of splines used to build up the fit line), and as such gene expression trends should generally be viewed as qualitative measures of gene expression changes. In MiCV, these trends are dynamically generated in both whole-dataset and branch-specific manners depending on user need using pyGAM (Servén et al., 2018).

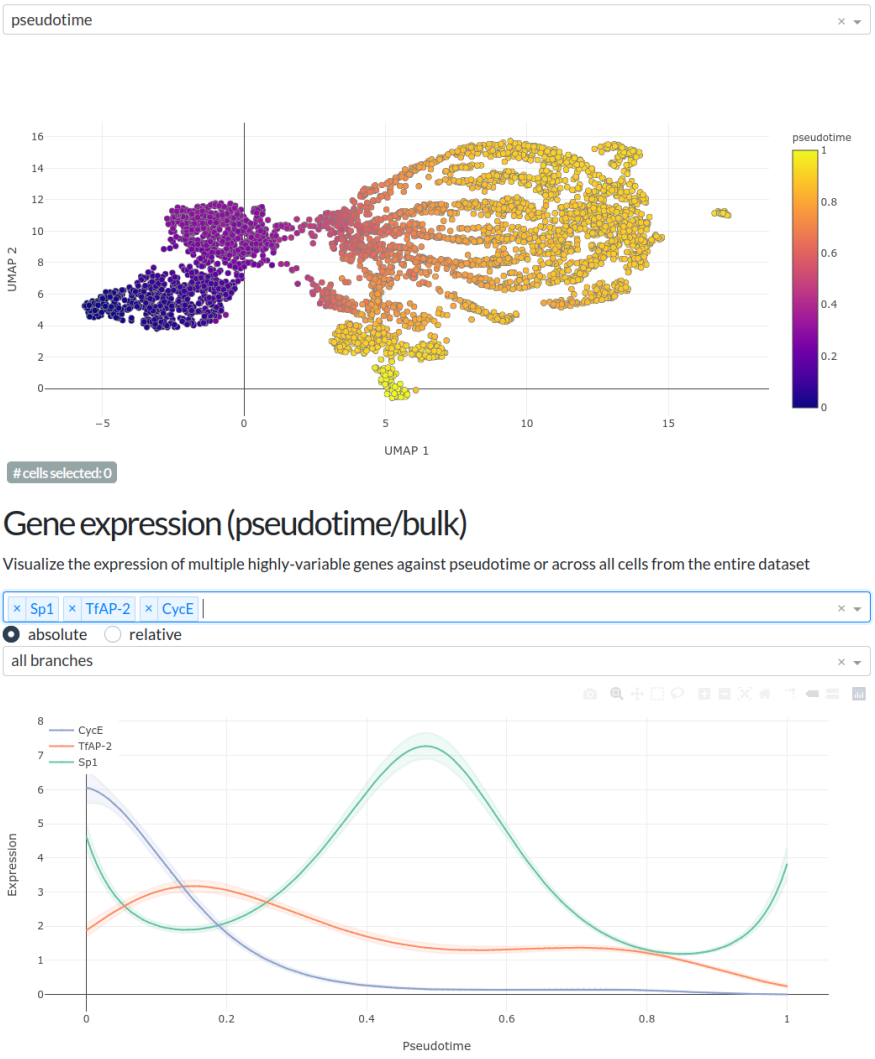

Fig. 3.19: Pseudotime analysis in MiCV
Top: Cells from the type II scRNA-seq dataset projected in UMAP space and colored by their pseudotime trajectory placements. Bottom: Gene expression trends along pseudotime for 3 genes (*Sp1, TfAP-2, Fas3*) in the type II scRNA-seq dataset. Peaks/troughs for each gene can be readily identified, indicating points in the

differentiation trajectory where expression for each gene is high/low. Though plotted here in terms of absolute

Faster iterations of iterative scRNA-seq secondary analysis

A great many scRNA-seq experiments capture a wide variety of cell types, each of which may be made up of many cellular subtypes. In order to facilitate probing this hierarchical cell type structure, researchers often adopt an iterative analysis paradigm, wherein they first do broad cell type clustering (and marker gene analysis), and then break each broad cell type out into its own dataset before repeating this analysis process on each broad subtype independently. MiCV facilitates this type of analysis in two ways. Firstly, MiCV provides an annotation tab/webpage, wherein researchers can use automated cell type annotations, gene expression plots, pseudotime trajectories, gene expression trends, and gene information from external databases to select cells of interest and extract them for analysis apart from the rest of the dataset. This is done non-destructively by adding user-provided annotations to the dataset and saving subsets of the data as independent copies that can be analyzed independently of one another. Secondly, MiCV provides a ‘summary report’ function that takes in an scRNA-seq dataset and performs clustering and iterative subclustering analysis automatically for a user, ending with a printable PDF summary report that includes cell type UMAP plots, marker gene analysis, database information, and iterative subclustering results. A relatively novel feature in the space of user-facing scRNA-seq secondary analysis applications, these summary reports facilitate the rapid viewing and sharing of automatic secondary analysis, making it possible for multiple researchers with no programming experience to collectively analyze a dataset. Though not interactive, these summary reports make it possible to rapidly inform the direction of iterative, interactive analysis steps by bringing more researchers and literature context together to tackle these complex datasets.

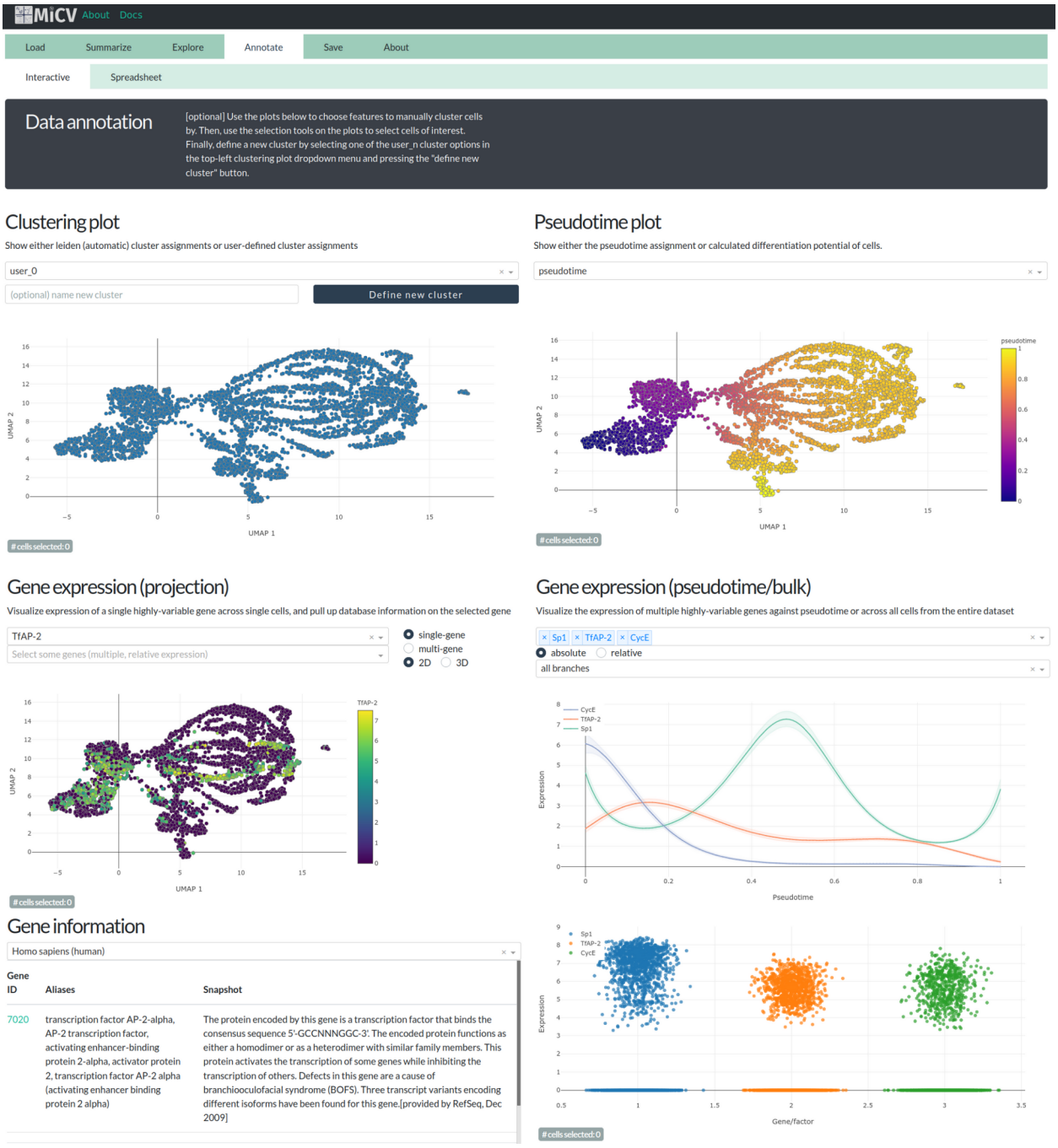

Fig. 3.20: Manually annotating cell types in MiCV

Integrating multiple bioinformatic analysis results, this page enables researchers to cross-filter cells by selecting them on any/all of the plots on this page, manually assigning cell types based on the researcher's desired set of criteria (in a similar fashion to cell gating in FACS software suites).

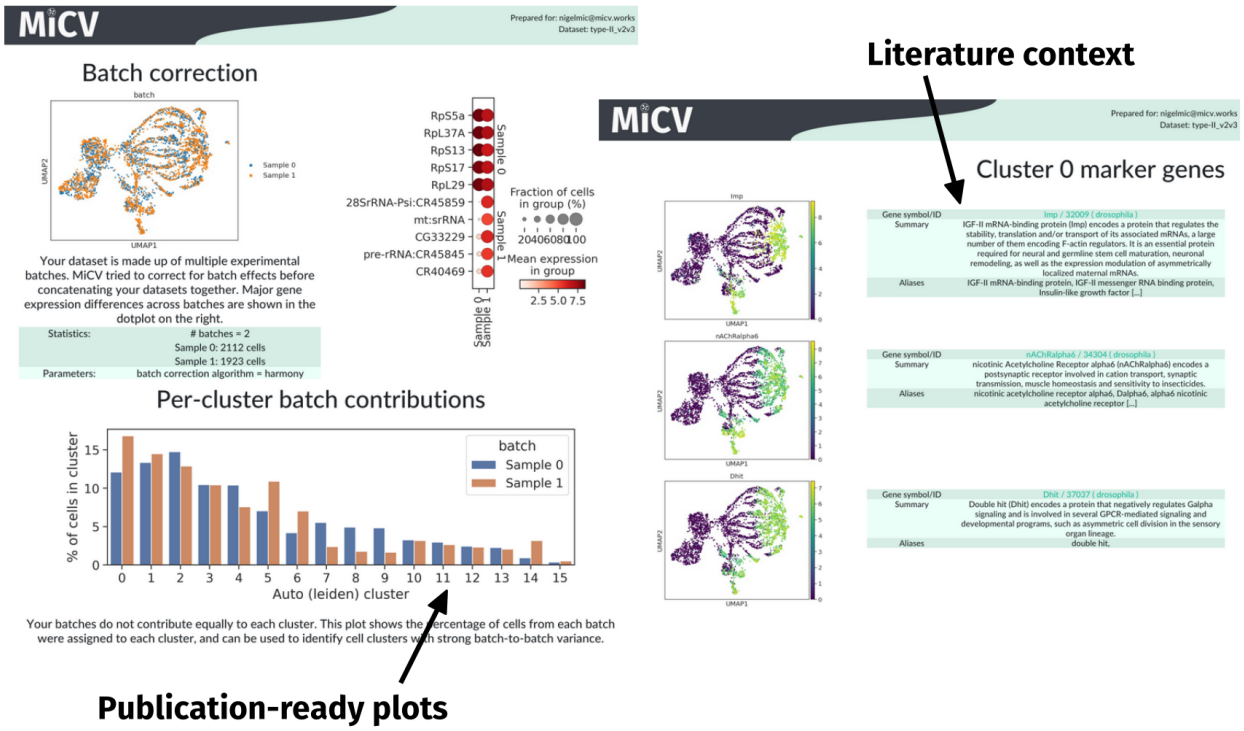

Fig. 3.21: The MiCV summary report
In a completely automated fashion, MiCV takes in scRNA-seq datasets, performs the default scRNA-seq secondary analysis pipeline, generates publication ready plots with plain-language descriptors, and provides additional literature context to accelerate time to discovery. These shareable PDF reports facilitate the inclusion of many people when analyzing and forming hypotheses based around large and complex scRNA-seq datasets.

The future of single-cell experimental scope, design, and data analysis

In this chapter I have provided a brief outline of the steps involved in profiling mRNA expression at the single-cell level, walking through sample preparation, cell sorting/isolation, mRNA capture, sequencing, and both primary mapping as well as secondary counts matrix analysis. This outline, and the development of MiCV as a whole, was largely centered around needs and challenges I identified around the previously described scRNA-seq experiments we performed in order to profile the type II NB system in *Drosophila*. However it is generally appreciated that scRNA-seq, and more broadly the field of single-cell -omics (transcriptomics, proteomics, epigenomics, etc.) has been *rapidly* expanding since the development of droplet-based scRNA-seq

methodologies in 2015 (Aldridge and Teichmann, 2020; Macosko et al., 2015), and continues to do so to this day. This growth can largely be attributed to advances on the experimental design and secondary analysis frontiers, both of which I will outline here as future directions for my work in the type II NB system of *Drosophila* as well as for the single-cell secondary analysis tool I have developed.

A central feature of all scRNA-seq experimental protocols is the capture and barcoding with unique cellular barcodes of mRNA from individual cells. From a biochemical perspective, developing ssDNA oligos with cell barcodes and poly-dT regions that can bind the poly-A tails of mRNA makes this possible; however, there is no reason why this capture need be limited to single-stranded oligos that contain long poly-A regions. As such, researchers have commandeered the use of droplet-based mRNA capture technologies to profile other aspects of cell status, including chromatin accessibility (scATAC-seq) (Buenrostro et al., 2015), DNA methylation (Mulqueen et al., 2018), and cell surface protein expression (Peterson et al., 2017) becoming both common and commercially supported. Additionally, work has been done to combine multiple assays into single reactions, enabling simultaneous readouts of both mRNA and surface protein expression in the same cells (Peterson et al., 2017), for example. Integrating this information to link chromatin accessibility, mRNA, and protein expression can make it possible to address complex hypotheses, such as “What is the link between the accessibility of this region of the genome and mRNA expression of genes in that locus - and when mRNA is finally made, what protein expression changes can we observe in this specific cell type? Is this mRNA degraded before protein can be expressed? Is this protein long-lived relative to its mRNA?”.

Furthermore, these same biochemistries have been translated from droplets to slides, enabling so-called ‘spatial sequencing’ assays (Asp et al., 2020; Cho et al., 2021; Liu et al., 2020; Rodriques et al., 2019). Typically performed by laying a fixed and permeabilized tissue slice atop a regular, micron-resolution grid of mRNA capture sites, these protocols generate not single-cell but single-micron resolution gene expression

data. In this way, specific regions of tissues can be profiled for thousands of genes in parallel, competing now with resources such as the Allen Brain Atlas (which profiles gene expression one gene at a time, compiling *in situ hybridization* results from the brains of thousands of animals) for utility (Lein et al., 2007). Alternative approaches to 'spatial sequencing' imaging tissues prior to dissection/dissociation, aiming to identify cells of interest (typically those that express genetically-encoded fluorescent proteins) and 'keep track' of them during the mRNA capture process, as in (Nichterwitz et al., 2016) and (Biase et al., 2018). Though currently throughput and sequencing quality limited, improvements to these techniques through the use of more robust multi-color barcoding and high-throughput imaging techniques such as bitbow (Li et al., 2020b; Veling et al., 2019) may yield promising results without the need for more costly tissue preps and spatial sequencing reagents.

These new techniques make it possible to address new hypotheses in a high-throughput, unbiased manner, but only when coupled with secondary analysis tools that can accurately model and transform these new data types. Some work has been done to adopt previously developed scRNA-seq analysis tools to accept these new data formats, especially in the scATAC-seq (Butler et al., 2018; Danese et al., 2019) and single-cell surface protein sequencing (CITE/REAP-seq) (Lopez et al., 2018; Svensson et al., 2020) spaces where a 'counts matrix' akin to that in scRNA-seq data is still generated; here representing which genes are accessible and which surface proteins are detected, respectively. Spatial transcriptomic datasets, more visual in nature, are beginning to see the development of graphical user interfaces that enable interactive exploration of these datasets in a way akin to traditional fluorescence microscopy analysis (Palla et al., 2021).

As these methods continue to develop, new computational approaches that improve analytical rigour, computational efficiency, and user accessibility will hopefully continue to develop alongside them. The strong dependency MiCV has on similar open source libraries will enable its continued development, making adding in support for these new

approaches trivial. It is my hope that by continuing to incorporate and provide intuitive interfaces for these new open source tools, MiCV will continue to accelerate our time to discovery in single-cell -omics projects as the field develops, further elucidating what makes each and every cell in the world unique.

References

Aldridge, S., and Teichmann, S.A. (2020). Single cell transcriptomics comes of age. Nat. Commun. *11*, 4307.

Alles, J., Karaiskos, N., Praktiknjo, S.D., Grosswendt, S., Wahle, P., Ruffault, P.-L., Ayoub, S., Schreyer, L., Boltengagen, A., Birchmeier, C., et al. (2017). Cell fixation and preservation for droplet-based single-cell transcriptomics. BMC Biol. *15*, 44.

Asp, M., Bergenstråhle, J., and Lundeberg, J. (2020). Spatially Resolved Transcriptomes-Next Generation Tools for Tissue Exploration. Bioessays *42*, e1900221.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Berger, C., Harzer, H., Burkard, T.R., Steinmann, J., van der Horst, S., Laurenson, A.-S., Novatchkova, M., Reichert, H., and Knoblich, J.A. (2012). FACS purification and transcriptome analysis of drosophila neural stem cells reveals a role for Klumpfuss in self-renewal. Cell Rep. *2*, 407–418.

Biase, F.H., Wu, Q., Calandrelli, R., Rivas-Astroza, M., Zhou, S., Chen, Z., and Zhong, S. (2018). Rainbow-Seq: Combining Cell Lineage Tracing with Single-Cell RNA Sequencing in Preimplantation Embryos. IScience *7*, 16–29.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Cao, Y., Kitanovski, S., Küppers, R., and Hoffmann, D. UMI or not UMI, that is the question for scRNA-seq zero-inflation. Nat. Biotechnol. *39*, 158–159.

Cho, C.-S., Xi, J., Park, S.-R., Hsu, J.-E., Kim, M., Jun, G., Kang, H.-M., and Lee, J.H. (2021). Seq-Scope: Submicrometer-resolution spatial transcriptomics for single cell and subcellular studies. BioRxiv.

Croset, V., Treiber, C.D., and Waddell, S. (2017). Cellular diversity in the Drosophila midbrain revealed by single-cell transcriptomics. BioRxiv.

Croset, V., Treiber, C.D., and Waddell, S. (2018). Cellular diversity in the *Drosophila* midbrain revealed by single-cell transcriptomics. Elife *7*.

Danese, A., Richter, M.L., Fischer, D.S., Theis, F.J., and Colomé-Tatché, M. (2019). EpiScanpy: integrated single-cell epigenomic analysis. BioRxiv.

Davie, K., Janssens, J., Koldere, D., De Waegeneer, M., Pech, U., Kreft, Ł., Aibar, S., Makhzami, S., Christiaens, V., Bravo González-Blas, C., et al. (2018). A Single-Cell Transcriptome Atlas of the Aging *Drosophila* Brain. Cell *174*, 982-998.e20.

Denisenko, E., Guo, B.B., Jones, M., Hou, R., de Kock, L., Lassmann, T., Poppe, D., Clément, O., Simmons, R.K., Lister, R., et al. (2020). Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. *21*, 130.

Dhapola, P., Rodhe, J., Olofzon, R., Bonald, T., Erlandsson, E., Soneji, S., and Karlsson, G. (2021). Scarf: A toolkit for memory efficient analysis of large-scale single-cell genomics data. BioRxiv.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell *174*, 716-729.e27.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Gilboa, E., Mitra, S.W., Goff, S., and Baltimore, D. (1979). A detailed model of reverse transcription and tests of crucial aspects. Cell *18*, 93–100.

Goldberger, J., Hinton, G.E., Roweis, S., and Salakhutdinov, R.R. (2004). Neighbourhood Components Analysis. Advances in Neural Information Processing Systems.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods *11*, 637–640.

Habib, N., Basu, A., Avraham-Davidi, I., Burks, T., Choudhury, S.R., Aguet, F., Gelfand, E., Ardlie, K., Weitz, D.A., Rozenblatt-Rosen, O., et al. (2017). DroNc-Seq: Deciphering cell types in human archived brain tissues by massively-parallel single nucleus RNA-seq. BioRxiv.

Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296.

Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A.J.M., Faridani, O.R., and Sandberg, R. (2020). Single-cell RNA counting at allele and isoform resolution using Smart-seq3. Nat. Biotechnol. *38*, 708–714.

Harzer, H., Berger, C., Conder, R., Schmauss, G., and Knoblich, J.A. (2013). FACS purification of Drosophila larval neuroblasts for next-generation sequencing. Nat. Protoc. *8*, 1088–1099.

Herzenberg, L.A., Parks, D., Sahaf, B., Perez, O., Roederer, M., and Herzenberg, L.A. (2002). The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. Clin. Chem. *48*, 1819–1827.

Hochgerner, H., Lönnerberg, P., Hodge, R., Mikes, J., Heskol, A., Hubschle, H., Lin, P., Picelli, S., La Manno, G., Ratz, M., et al. (2017). STRT-seq-2i: dual-index 5' single cell and nucleus RNA-seq on an addressable microwell array. Sci. Rep. *7*, 16327.

Joglekar, A., Prjibelski, A., Mahfouz, A., Collier, P., Lin, S., Schlusche, A.K., Marrocco, J., Williams, S.R., Haase, B., Hayes, A., et al. (2021). A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. Nat. Commun. *12*, 463.

Kaminow, B., Yunusov, D., and Dobin, A. (2021). STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. BioRxiv.

Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H.B., et al. (2020). CellRank for directed single-cell fate mapping. BioRxiv.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature *445*, 168–176.

Lin, E., Rivera-Báez, L., Fouladdel, S., Yoon, H.J., Guthrie, S., Wieger, J., Deol, Y., Keller, E., Sahai, V., Simeone, D.M., et al. (2017). High-Throughput Microfluidic Labyrinth for the Label-free Isolation of Circulating Tumor Cells. Cell Syst. *5*, 295-304.e4.

Liu, Y., Yang, M., Deng, Y., Su, G., Enninful, A., Guo, C.C., Tebaldi, T., Zhang, D., Kim, D., Bai, Z., et al. (2020). High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. Cell *183*, 1665-1681.e18.

Li, B., Gould, J., Yang, Y., Sarkizova, S., Tabaka, M., Ashenberg, O., Rosen, Y., Slyper, M., Kowalczyk, M.S., Villani, A.-C., et al. (2020a). Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. Nat. Methods *17*, 793–798.

Li, H., Horns, F., Wu, B., Xie, Q., Li, J., Li, T., Luginbuhl, D.J., Quake, S.R., and Luo, L.

(2017). Classifying Drosophila Olfactory Projection Neuron Subtypes by Single-Cell RNA Sequencing. Cell *171*, 1206-1220.e22.

Li, Y., Walker, L.A., Zhao, Y., Edwards, E.M., Michki, N.S., Cheng, H.P.J., Ghazzi, M., Chen, T.Y., Chen, M., Roossien, D.H., et al. (2020b). Bitbow: a digital format of Brainbow enables highly efficient neuronal lineage tracing and morphology reconstruction in single brains. BioRxiv.

Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T., participants in the 1st Human Cell Atlas Jamboree, and Marioni, J.C. (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. *20*, 63.

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. *33*, D54-8.

McCarthy, D.J., Campbell, K.R., Lun, A.T.L., and Wills, Q.F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics *33*, 1179–1186.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv:1802.03426 [Cs, Stat].

Melsted, P., Booeshaghi, A.S., Liu, L., Gao, F., Lu, L., Min, K.H.J., da Veiga Beltrame, E., Hjörleifsson, K.E., Gehring, J., and Pachter, L. (2021). Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nat. Biotechnol.

Michki, N.S., Li, Y., Sanjasaz, K., Zhao, Y., Shen, F.Y., Walker, L.A., Cao, W., Lee, C.-Y., and Cai, D. (2021). The molecular landscape of neural differentiation in the developing Drosophila brain revealed by targeted scRNA-seq and multi-informatic analysis. Cell Rep. *35*, 109039.

Miltenyi, S., Müller, W., Weichel, W., and Radbruch, A. (1990). High gradient magnetic cell separation with MACS. Cytometry *11*, 231–238.

Mulqueen, R.M., Pokholok, D., Norberg, S.J., Torkenczy, K.A., Fields, A.J., Sun, D., Sinnamon, J.R., Shendure, J., Trapnell, C., O'Roak, B.J., et al. (2018). Highly scalable generation of DNA methylation profiles in single cells. Nat. Biotechnol. *36*, 428–431.

Nichterwitz, S., Chen, G., Aguila Benitez, J., Yilmaz, M., Storvall, H., Cao, M., Sandberg, R., Deng, Q., and Hedlund, E. (2016). Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. Nat. Commun. *7*, 12139.

Ntranos, V., Yi, L., Melsted, P., and Pachter, L. (2018). Identification of transcriptional signatures for cell types from single-cell RNA-Seq. BioRxiv.

Palla, G., Spitzer, H., Klein, M., Fischer, D.S., Schaar, A.C., Kuemmerle, L.B., Rybakov, S., Ibarra, I.L., Holmberg, O., Virshup, I., et al. (2021). Squidpy: a scalable framework for spatial single cell analysis. BioRxiv.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6 *2*, 559–572.

Peterson, V.M., Zhang, K.X., Kumar, N., Wong, J., Li, L., Wilson, D.C., Moore, R., McClanahan, T.K., Sadekova, S., and Klappenbach, J.A. (2017). Multiplexed quantification of proteins and transcripts in single cells. Nat. Biotechnol. *35*, 936–939.

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Pretlow, T., and Pretlow, T. (1987). Cell Separation (Academic Press).

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979–982.

Richardson, G.M., Lannigan, J., and Macara, I.G. (2015). Does FACS perturb gene expression? Cytometry A *87*, 166–175.

Rizzardi, L.F., Kunz, H., Rubins, K., Chouker, A., Quiriarte, H., Sams, C., Crucian, B.E., and Feinberg, A.P. (2016). Evaluation of techniques for performing cellular isolation and preservation during microgravity conditions. NPJ Microgravity *2*, 16025.

Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science *363*, 1463–1467.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Servén, D., Brummitt, C., and Abedi, H. (2018). pyGAM: Generalized Additive Models in Python (Zenodo).

Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'er, D. (2019). Characterization of cell fate probabilities in single-cell data with Palantir. Nat. Biotechnol. *37*, 451–460.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science *309*, 1728–1732.

Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. *27*, 491–499.

Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics *19*, 477.

Svensson, V. (2019). Droplet scRNA-seq is not zero-inflated. BioRxiv.

Svensson, V., Gayoso, A., Yosef, N., and Pachter, L. (2020). Interpretable factor models of single-cell RNA-seq via variational autoencoders. Bioinformatics *36*, 3418–3421.

Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233.

Tran, T.N., and Bader, G. (2019). Tempora: cell trajectory inference using time-series single-cell RNA sequencing data. BioRxiv.

Veling, M.W., Li, Y., Veling, M.T., Litts, C., Michki, N., Liu, H., Ye, B., and Cai, D. (2019). Identification of Neuronal Lineages in the Drosophila Peripheral Nervous System with a "Digital" Multi-spectral Lineage Tracing System. Cell Rep. *29*, 3303-3312.e3.

Vuong, H., Truong, T., Phan, T., and Pham, S. (2020). Venice: A new algorithm for finding marker genes in single-cell transcriptomic data. BioRxiv.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M.,

Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol. Cell *65*, 631-643.e4.

Chapter 4

Conclusions and Future Directions

Summary of type-II neurogenesis work in *Drosophila melanogaster*

Drosophila melanogaster represents a model organism that recapitulates many features of vertebrate neurogenesis. Unlike the abundant type-I neuroblasts (NB, neural stem cells), the 16 type II NBs in the *Drosophila* brain adopt a neurogenesis process that is directly analogous to that observed in mammalian cortical development (Homem and Knoblich, 2012). During development, each type II NB undergoes repeated asymmetric cell divisions to generate an NB and a sibling progeny that acquires a progenitor identity (i.e. intermediate neural progenitor, INP). Each INP undergoes limited rounds of asymmetric cell division to re-generate and to produce a ganglion mother cell (GMC), which divides once more to become two neuron(s) and/or glial cell(s). Along this NB-INP-GMC-neuron maturation process, cells express a well-defined cascade of transcription factors that mark these cell differentiation stages (Ren et al., 2017; Syed et al., 2017). In parallel, INPs born in each division cycle may express a cascade of transcription factors unique to each NB lineage that contribute to the generation of different neural progenies (Bayraktar and Doe, 2013). It is highly plausible that the combination of these two transcription factor cascades alongside a third molecular axis, which defines unique NBs (i.e., each NB generates a distinct lineage), brings about the generation of a highly diverse neuronal pool.

In chapter 1, I described the current state of the field surrounding type-II neurogenesis in *Drosophila*. In particular, I laid out how temporally varying transcription factor cascades modulate the neurogenic potential of type-II NBs and INPs, and discuss low-

throughput antibody/functional manipulation studies have previously been used to characterize these changes during neural development. Despite the massive significance of these studies, their reliance on antibody libraries limits their scope. *Drosophila melanogaster* has a genome with more than 14,000 protein-coding genes, at least 700 of which exhibit transcription factor activities (Shokri et al., 2019). Though direct-screening techniques are powerful, the advent of high throughput single-cell mRNA sequencing (scRNA-seq) technologies has enabled researchers to much more broadly investigate the mRNA expression landscape of hundreds of thousands of cells (Macosko et al., 2015; Ziegenhain et al., 2017). Coupled with a vast array of analytical tools that enable us to take these high-dimensional datasets and identify significant molecular signatures from them (Butler et al., 2018; Wolf et al., 2018), researchers can make hypotheses about the number of unique cellular subtypes in the brain (Cocanougher et al., 2019; Saunders et al., 2018), what the functions of these subtypes might be (Ren et al., 2019), and what subtypes might arise together along a common developmental pathway (Cao et al., 2019; Qiu et al., 2017; Soldatov et al., 2019). This combination of experimental and analytical advances has removed the need to rely solely on limited and biased single-molecule screens, instead enabling experiments to yield information about the near-complete mRNA expression landscape of thousands of cells at a time without the need for prior knowledge about/reagents for probing specific genes of interest.

In chapter 2, I described how we used targeted single cell transcriptome analysis to advance our understanding of the *Drosophila* type-II neuron differentiation process. After initially separating the transcriptomes of the type-II neuroblast derived cells from those labeled in the optic lobes, we show that pseudotime analysis techniques can be used to define a maturation axis and extract putative marker genes that specify the INP, GMC, immature neuron and mature neuron differentiation stages. Broadly expressed, not limited to the type-II NB progenies, these marker genes of different maturation stages indeed form intersectional patterns that represent the spatial organization of the neurogenesis progress in the larval brain. Compared to previous antibody-based and

gene manipulation-based screening strategies, scRNA-seq data permits a high-throughput assessment of the whole gene expression profile to rapidly identify candidate genes for functional study. For instance, in the past, Hey has been shown to mark one of the two immature neurons derived from the final cell division, and its role is exclusive as an inhibitor of Notch signaling in this immature neuron (Monastirioti et al., 2010). From our scRNAseq analysis, E(spl)m6-BFM, a member of the enhancer-of-split family of transcription factors (Lai et al., 2000), and Rbp, a rim-binding protein responsible for synaptic homeostasis and neurotransmitter release (Liu et al., 2011; Müller et al., 2015) are exclusively up-regulated in only the transient immature neuronal differentiation state directly after GMC division. These two marker genes can be used to guide the exploration of Hey- immature neurons in future studies. Functional knock-outs of these two genes will be critical to understanding their function in newly-born neurons as it pertains to their maturation and any early functional role they may play in the developing brain.

Further higher-resolution clustering of the INP and GMC cells identified transcriptomically correlated subclusters between these two stages, which supports the idea that parallel maturation transitions happen at the same developmental time point. However, scRNA-seq data alone cannot distinguish whether these parallel transitions are due to the co-existence of earlier and newly born INPs in all NB lineages or due to the intrinsic differences among NB lineages. We therefore in situ profiled the marker genes selected from the scRNA-seq selected candidates and restored their missing spatial information that indicates the maturation stage as well as the NB lineage identity. In addition, combined with prior knowledge, whether a marker gene is expressed in younger or earlier born INPs can also be speculated. Our findings conclude that Sp1 is expressed in the young INPs of nearly all NB lineages, whereas TfAP-2 and Fas3 express in older INPs belonging to specific NB lineages. Interestingly, we found that Sp1 and TfAP-2 expressed not only in neural progenitors but also in maturing neurons. These transcription factors seem to intermingle with the NB lineage-specific D/grh/ey cascades in the INP stage, but eventually differentiate into completely exclusive neuron

populations. Finally, higher-resolution clustering of neurons in our scRNAseq dataset revealed that transcription factors and surface molecules are predominant markers for distinct neuronal subtypes at the 3rd instar larval stage. This implies that most neurons of the type-II NB progenies have not started to gain their differentiated functions at this stage of development.

Combining in silico scRNA-seq analysis and in situ mRNA imaging, we discovered many transcription factors and surface molecules that potentially play important roles in generating neuronal subtypes in an NB-specific, INP-specific, or function-specific manner. These discoveries helped us to gain a comprehensive understanding of the molecular landscape along all three major neural developmental axes that define a cell's progenitor lineage identity, progenitor cell division number, and differentiation state (Fig. 2.14). This model provides a general guidance for biologists to disentangle the differentiation process in complex systems beyond the Drosophila brain.

Though the scope of our work here may have been considered unprecedented 6 years ago (scope here referring to how we sequenced approximately 4000 cells that were neurons originating from 16 Drosophila type-II neuroblast lineages), many open questions still remain. With low-resolution clustering, we identified 13 molecularly distinct neural subtypes. Increasing the clustering resolution just a bit higher we could identify more than 20 that are still distinct (data not shown). Similarly, as we show with the INPs/GMCs in our dataset, a low-resolution clustering can often mask the cellular diversity that is present in the system. As we know that each type-II neuroblast generates approximately 38 INPs throughout their developmental lifespan (Bayraktar et al., 2010; Bello et al., 2008), the presented clustering in this paper only captures part of the INP diversity. One straightforward thought is to increase the number of sequenced single cells so that higher clustering resolution may eventually reveal even the most subtle differences between each of the hundreds of INPs in the type-II system. However, as transcription factor cascades involved in INP division/maturation intertwine with those involved in NB specification and differentiation, we expect that the INP heterogeneity can be untangled somewhat using a higher clustering resolution but still

fails to provide us with a coherent view of the complex lineage, maturation, and differentiation landscape we are attempting to characterize. These issues highlight the challenge of deconvoluting the INP maturation, NB lineage, and differentiation state axes and the need for a holistic, integrated approach to experimental design and subsequent bioinformatic analysis.

The data we have presented here were collected at a single developmental time-point (late third instar), but we know that type-II neurogenesis precedes and continues after this stage. Repeating these scRNA-seq experiments at more developmental time-points will reveal more in what order molecularly-defined neural subsets are generated. Using recently developed analytical techniques to “stitch” these multi-time-point datasets together (Lin et al., 2019; Tran and Bader, 2019) will be advantageous to align all the cells along a unified developmental time axis. To overcome the limitation of the R9D11-Gal4 driver, which does not label neuroblasts nor the fully mature neurons, a permanent labeling strategy, similar to the one used in (Bayraktar et al., 2010) but covering all lineages more reliably for FACS, is required. More critically, such permanent labeling needs to be paired with technologies that provide single-lineage specification resolution, such as the introduction of single-neuroblast lineage barcoding techniques. Genetic constructs based around CRISPR-Cas9 (Raj et al., 2017; Spanjaard et al., 2018) and the Cre/Lox system (Kalhor et al., 2018; Pei et al., 2017; Weber et al., 2016) have been developed for this purpose, although which exact lineage was labeled by a particular barcode was still unknown. The introduction of a spectrally unique barcode for each neuroblast lineage, in a similar vein to the recently developed Bitbow lineage tracking strategy (Li et al., 2020; Veling et al., 2019), would be advantageous as they can provide direct in situ evidence for neuroblast lineage identity.

Finally, our work identifies several transcription factors that are specifically expressed in subsets of cells of the type-II neuroblast progenies. Our in silico and in situ results showed that their expressions are either constrained to particular developmental stages, or in subsets of cells that are born in different orders. It would be desired to perform

follow up experiments to reveal whether these transcription factors play important roles in specifying the terminal fates of type-II neuronal subtypes.

The future of single-cell experimental scope, design, and data analysis

In chapter 3 I have provided a brief outline of the steps involved in profiling mRNA expression at the single-cell level, walking through sample preparation, cell sorting/isolation, mRNA capture, sequencing, and both primary mapping as well as secondary counts matrix analysis. This outline, and the development of MiCV as a whole, was largely centered around needs and challenges I identified around the previously described scRNA-seq experiments we performed in order to profile the type II NB system in *Drosophila*. However it is generally appreciated that scRNA-seq, and more broadly the field of single-cell -omics (transcriptomics, proteomics, epigenomics, etc.) has been *rapidly* expanding since the development of droplet-based scRNA-seq methodologies in 2015 (Aldridge and Teichmann, 2020; Macosko et al., 2015), and continues to do so to this day. This growth can largely be attributed to advances on the experimental design and secondary analysis frontiers, both of which I will outline here as future directions for my work in the type II NB system of *Drosophila* as well as for the single-cell secondary analysis tool I have developed.

A central feature of all scRNA-seq experimental protocols is the capture and barcoding with unique cellular barcodes of mRNA from individual cells. From a biochemical perspective, developing ssDNA oligos with cell barcodes and poly-dT regions that can bind the poly-A tails of mRNA makes this possible; however, there is no reason why this capture need be limited to single-stranded oligos that contain long poly-A regions. As such, researchers have commandeered the use of droplet-based mRNA capture technologies to profile other aspects of cell status, including chromatin accessibility (scATAC-seq) (Buenrostro et al., 2015), DNA methylation (Mulqueen et al., 2018), and cell surface protein expression (Peterson et al., 2017) becoming both common and commercially supported. Additionally, work has been done to combine multiple assays

into single reactions, enabling simultaneous readouts of both mRNA and surface protein expression in the same cells (Peterson et al., 2017), for example. Integrating this information to link chromatin accessibility, mRNA, and protein expression can make it possible to address complex hypotheses, such as “What is the link between the accessibility of this region of the genome and mRNA expression of genes in that locus - and when mRNA is finally made, what protein expression changes can we observe in this specific cell type? Is this mRNA degraded before protein can be expressed? Is this protein long-lived relative to its mRNA?”.

Furthermore, these same biochemistries have been translated from droplets to slides, enabling so-called ‘spatial sequencing’ assays (Asp et al., 2020; Cho et al., 2021; Liu et al., 2020; Rodriques et al., 2019). Typically performed by laying a fixed and permeabilized tissue slice atop a regular, micron-resolution grid of mRNA capture sites, these protocols generate not single-cell but single-micron resolution gene expression data. In this way, specific regions of tissues can be profiled for thousands of genes in parallel, competing now with resources such as the Allen Brain Atlas (which profiles gene expression one gene at a time, compiling *in situ hybridization* results from the brains of thousands of animals) for utility (Lein et al., 2007). Alternative approaches to ‘spatial sequencing’ imaging tissues prior to dissection/dissociation, aiming to identify cells of interest (typically those that express genetically-encoded fluorescent proteins) and ‘keep track’ of them during the mRNA capture process, as in (Nichterwitz et al., 2016) and (Biase et al., 2018). Though currently throughput and sequencing quality limited, improvements to these techniques through the use of more robust multi-color barcoding and high-throughput imaging techniques such as bitbow (Li et al., 2020b; Veling et al., 2019) may yield promising results without the need for more costly tissue preps and spatial sequencing reagents.

These new techniques make it possible to address new hypotheses in a high-throughput, unbiased manner, but only when coupled with secondary analysis tools that can accurately model and transform these new data types. Some work has been done

to adopt previously developed scRNA-seq analysis tools to accept these new data formats, especially in the scATAC-seq (Butler et al., 2018; Danese et al., 2019) and single-cell surface protein sequencing (CITE/REAP-seq) (Lopez et al., 2018; Svensson et al., 2020) spaces where a ‘counts matrix’ akin to that in scRNA-seq data is still generated; here representing which genes are accessible and which surface proteins are detected, respectively. Spatial transcriptomic datasets, more visual in nature, are beginning to see the development of graphical user interfaces that enable interactive exploration of these datasets in a way akin to traditional fluorescence microscopy analysis (Palla et al., 2021).

As these methods continue to develop, new computational approaches that improve analytical rigour, computational efficiency, and user accessibility will hopefully continue to develop alongside them. The strong dependency MiCV has on similar open source libraries will enable its continued development, making adding in support for these new approaches trivial. It is my hope that by continuing to incorporate and provide intuitive interfaces for these new open source tools, MiCV will continue to accelerate our time to discovery in single-cell -omics projects as the field develops, further elucidating what makes each and every cell in the world unique.

References

Aldridge, S., and Teichmann, S.A. (2020). Single cell transcriptomics comes of age. Nat. Commun. 11, 4307.

Asp, M., Bergenstråhle, J., and Lundeberg, J. (2020). Spatially Resolved Transcriptomes-Next Generation Tools for Tissue Exploration. Bioessays 42, e1900221.

Biase, F.H., Wu, Q., Calandrelli, R., Rivas-Astroza, M., Zhou, S., Chen, Z., and Zhong, S. (2018). Rainbow-Seq: Combining Cell Lineage Tracing with Single-Cell RNA Sequencing in Preimplantation Embryos. IScience 7, 16–29.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. 36, 411–420.

Cho, C.-S., Xi, J., Park, S.-R., Hsu, J.-E., Kim, M., Jun, G., Kang, H.-M., and Lee, J.H. (2021). Seq-Scope: Submicrometer-resolution spatial transcriptomics for single cell and subcellular studies. BioRxiv.

Danese, A., Richter, M.L., Fischer, D.S., Theis, F.J., and Colomé-Tatché, M. (2019). EpiScanpy: integrated single-cell epigenomic analysis. BioRxiv.

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature 445, 168–176.

Liu, Y., Yang, M., Deng, Y., Su, G., Enninful, A., Guo, C.C., Tebaldi, T., Zhang, D., Kim, D., Bai, Z., et al. (2020). High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. Cell 183, 1665-1681.e18.

Li, Y., Walker, L.A., Zhao, Y., Edwards, E.M., Michki, N.S., Cheng, H.P.J., Ghazzi, M., Chen, T.Y., Chen, M., Roossien, D.H., et al. (2020). Bitbow: a digital format of Brainbow enables highly efficient neuronal lineage tracing and morphology reconstruction in single brains. BioRxiv.

Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods 15, 1053–1058.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214.

Mulqueen, R.M., Pokholok, D., Norberg, S.J., Torkenczy, K.A., Fields, A.J., Sun, D.,

Sinnamon, J.R., Shendure, J., Trapnell, C., O'Roak, B.J., et al. (2018). Highly scalable generation of DNA methylation profiles in single cells. Nat. Biotechnol. 36, 428–431.

Nichterwitz, S., Chen, G., Aguila Benitez, J., Yilmaz, M., Storvall, H., Cao, M., Sandberg, R., Deng, Q., and Hedlund, E. (2016). Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. Nat. Commun. 7, 12139.

Palla, G., Spitzer, H., Klein, M., Fischer, D.S., Schaar, A.C., Kuemmerle, L.B., Rybakov, S., Ibarra, I.L., Holmberg, O., Virshup, I., et al. (2021). Squidpy: a scalable framework for spatial single cell analysis. BioRxiv.

Peterson, V.M., Zhang, K.X., Kumar, N., Wong, J., Li, L., Wilson, D.C., Moore, R., McClanahan, T.K., Sadekova, S., and Klappenbach, J.A. (2017). Multiplexed quantification of proteins and transcripts in single cells. Nat. Biotechnol. 35, 936–939.

Ren, Q., Yang, C.-P., Liu, Z., Sugino, K., Mok, K., He, Y., Ito, M., Nern, A., Otsuna, H., and Lee, T. (2017). Stem Cell-Intrinsic, Seven-up-Triggered Temporal Factor Gradients Diversify Intermediate Neural Progenitors. Curr. Biol. 27, 1303–1313.

Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science 363, 1463–1467.

Shokri, L., Inukai, S., Hafner, A., Weinand, K., Hens, K., Vedenko, A., Gisselbrecht, S.S., Dainese, R., Bischof, J., Furger, E., et al. (2019). A Comprehensive Drosophila melanogaster Transcription Factor Interactome. Cell Rep. 27, 955-970.e7.

Svensson, V., Gayoso, A., Yosef, N., and Pachter, L. (2020). Interpretable factor models of single-cell RNA-seq via variational autoencoders. Bioinformatics 36, 3418–3421.

Syed, M.H., Mark, B., and Doe, C.Q. (2017). Steroid hormone induction of temporal gene expression in Drosophila brain neuroblasts generates neuronal and glial diversity. Elife 6.

Veling, M.W., Li, Y., Veling, M.T., Litts, C., Michki, N., Liu, H., Ye, B., and Cai, D. (2019). Identification of Neuronal Lineages in the Drosophila Peripheral Nervous System with a "Digital" Multi-spectral Lineage Tracing System. Cell Rep. 29, 3303-3312.e3.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19, 15.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol. Cell 65, 631-643.e4.

Appendices

A1: How Complex is Neural Fate Patterning? Theoretical Limits on Neural Diversity

In chapter 1 I described current literature around neural fate patterning mechanisms, with the following major mechanisms playing a role in defining neural fate:

A. Neural stem cell lineage identity
B. Neural stem cell age/division number (global developmental time)
C. Intermediate neural progenitor age/cell division number
D. Notch on/off state (when a precursor cell divides, one daughter cell is typically Notch on, the other Notch off)
E. Spatial targeting cues that direct neurons to make synaptic connections with specific neurons (can have retrograde effect on neural identity)

By assuming each of these mechanisms is discrete in nature and subsequently estimating the number of unique states each of these mechanisms can be in, we can begin to put limits on the number of neural fates possibly generated using these mechanisms. This assumption fails for facet E (the effect of spatial targeting cues), which is a more continuous, complex, and often subtle patterning mechanism, and so we will ignore it here.

Given the number of identifiable discrete states for each of the four remaining mechanisms, we can estimate the maximal number of unique neural fates as:

$$N = a \cdot b \cdot c \cdot d \quad \text{(eq. A1.1)}$$

where *a* is the number of discrete states for mechanism A (lineage identity), *b* the

number of for mechanism B (neural stem cell age), and likewise for *c* and *d*.

Based on previous work in the type II neuroblast system in *Drosophila*, and assuming we are interested in characterizing the number of unique neural fates up to the late 3rd instar larval stage (the context of this thesis), is is possible to estimate values for each of the variables in this equation:

A. There are 16 type II neuroblasts (8 per brain lobe), therefore *a=16*
B. It is not known precisely how many times each type II NB divides, therefore on the surface *b* is *unknown*. However, given that at the late L3 stage there are approximately 160 INPs total (80 INPs per brain lobe (Bayraktar et. al., 2010)) we can assume that each NB divides 10 times by late L3, and therefore *b=10*
C. Each INP divides between 4-8 times (an average of 6), therefore *c=6*. Knowing the median number of INP divisions would enable a better estimate for *c*.
D. Each precursor (GMC) divides once to generate one *Notch* on and one *Notch* off daughter cell, therefore *d=2*

Our estimate of the maximum number of unique neural fates then becomes:

$$N = ¿a \cdot b \cdot c \cdot d = (16) \cdot (10) \cdot (6) \cdot (2) = 1920 \quad \text{(eq. A1.2)}$$

which is in line with our total type II progeny cell count in the late L3 brain, based on images from the R9D11-Gal4 fly driver line crossed to our UAS-hH2b::2xmNG line (chapter 2). And indeed, we should expect the upper limit on the number of unique neural fates to be equal to the total number of neurons generated by the combination of these mechanisms - with these progenitor patterning-based fate specification mechanisms, we cannot have more neural fates than neurons to assign them to!

However, Eq. A1.2 represents a maxima that is potentially far greater than the *true* number of achievable neural fates. For it to be an accurate estimate of the true number of unique fates, a second assumption must be made about the independence of each of

these 4 mechanisms - namely, that they are completely independent of one another. This assumption likely does not hold, though the severity of the interdependence of these four mechanisms is challenging to estimate.

In Chapter 2, I determined experimentally that there are 13 obviously identifiable neural subtypes in the type II system at the late L3 stage of larval development - 20 subtypes, if we increase scRNA-seq clustering resolution. Since clustering resolution can be somewhat arbitrary, let us take 20 as our 'experimentally determined' number of unique neural fates - defined at least at the mRNA expression level. It is therefore apparent that:

$$N(maximum) \gg N(observed) \text{(eq. A1.3)}$$

since *N(maximum)=1920* and *N(observed)=20*, a difference of more than 2 orders of magnitude. While we expect our number of observed neural fates to be smaller than our estimated maximum, the scale of this difference warrants a refinement of our assumptions about each mechanism.

An important feature of brain development observed in almost all organisms with central nervous systems is lateral symmetry (brain lateralization), wherein across the two semi-independent lateral brain lobes there is obvious symmetry of gross anatomical features (Duboc et al. 2015). Though these two brain lobes are not completely identical, it stands to reason that in the relatively simple *Drosophila* developing CNS we can assume that the 16 type II NBs are laterally symmetric, with 8 in each lobe. Therefore:

$$a=8 \rightarrow N(maximum)=960 \text{ (eq. A1.4)}$$

We can continue to refine our estimate for *a* by incorporating what we learned from our *in situ* and scRNA-seq data related to INP subtypes; specifically, that the INPs from DM1, 2, and 3 all appear to express *Fas3*, and that INPs from DM4, 5, and 6 all appear

to express *TfAP-2*. Though these lineages are likely not *completely* identical (and indeed, previous work shows they are not (Bayraktar and Doe 2013), it stands to reason that there may be some redundancy among lineages DM1-3 and DM4-6, respectively. If we make the gross over-assumption than DM1-3 are replicated lineage, and likewise with DM4-6, then we have:

$$a=4 \rightarrow N(maximum)=480 \text{ (eq. A1.5)}$$

Additionally, we can refine our estimate for *c*, the number of neural fates we can attribute to INP patterning transitions. Initially we estimated this as the average number of times an INP divides (*c=6*), however the exemplary work by Bayraktar and Doe describing INP combinatorial patterning points in general towards each INP transitioning to express 3 unique transcription factors, with variability across NB lineages in the way these transcription factors co-express to define a unique patterning state (Bayraktar and Doe 2013). At the very least, they identify 4 uniquely identifiable neural fates arising from the DM 2-3 NB lineages. As such, we could arrive at:

$$c=4 \rightarrow N(maximum)=320 \text{ (eq. A1.6)}$$

The inclusion of additional mechanisms (for example E, the connectomic/retrograde signalling neural fate patterning mechanism) would only increase our estimate further. Taken together then, we might conclude that while this combinatorial patterning system makes *theoretically possible* more than 2000 unique neural fates, only approximately 1-5% of that potential 'fate-space' appears to *actually* be utilized by the late L3 stage of brain development in *Drosophila*.

Critically, this conclusion is dependent on a number of assumptions, chief among them that the 20 neural subtypes identified in scRNA-seq analysis are truly the only unique neural fates attainable by type II progeny at this stage. This number is likely an *underestimate* of the true neural diversity, due to the fact that scRNA-seq experiments

are currently only capable of sampling 1-10% of the total mRNA in a given mammalian cell (Shapiro et al. 2013), and that it is statistically unlikely that each unique neural fate is represented by a cell in our dataset (i.e. we 'missed' some cells in the single-cell sequencing process), though that we have sampled the vast majority of them is likely the case. Technological improvements in mRNA capture techniques, as well as improvements to genetic labelling and cell sorting techniques, will improve our resolution and neural subtype sampling completeness. Taken together, such improvements would be likely to *increase* the number of neural subtypes we would identify at any given developmental stage, should these experiments be repeated in the Be that as it may, these estimates combined with our scRNA-seq and *in situ* data point towards the conclusion that the type II system's 'goal', and thus the 'goal' of neurogenesis via intermediate neural progenitors in *Drosophila*, is not solely to maximize neural diversity. It is possible that robustness and redundancy in neural fate determination (i.e. ensuring that each fate is represented in the proper ratios despite potential cell death/failure to differentiate/etc.) is an additional benefit to using neurogenesis via intermediate progenitors. In conclusion, both neural diversity and robustness of neural pool generation may simultaneously be advanced by utilizing this mechanism, and further studies across this and other organisms may further refine the balance between these two benefits and show why neurogenesis via intermediate progenitors eventually became the predominant neurogenesis mechanism is higher vertebrates. future.

Be that as it may, these estimates combined with our scRNA-seq and *in situ* data point towards the conclusion that the type II system's 'goal', and thus the 'goal' of neurogenesis via intermediate neural progenitors in *Drosophila*, is not solely to maximize neural diversity. It is possible that robustness and redundancy in neural fate determination (i.e. ensuring that each fate is represented in the proper ratios despite potential cell death/failure to differentiate/etc.) is an additional benefit to using neurogenesis via intermediate progenitors. In conclusion, both neural diversity and robustness of neural pool generation may simultaneously be advanced by utilizing this

mechanism, and further studies across this and other organisms may further refine the balance between these two benefits and show why neurogenesis via intermediate progenitors eventually became the predominant mechanism in vertebrates.

A2: Reagents Used in Experimental Characterization of Type-II NB System

Table 2.2: Reagents categorized by type

Reagent	SOURCE	IDENTIFIER
Antibodies		
Mouse monoclonal anti-Fas3	DHSB	Cat# 7G10, RRID:AB_528238
Rat monoclonal anti-Dpn	Lee, C.Y., Robinson, K.J., and Doe, C.Q. (Nature, 2006a)	NA
Donkey-anti-Ms (AF488)	Jackson ImmunoResearch Laboratories, Inc.	Cat# 715-545-151
Donkey-anti-Rt (AF647)	Jackson ImmunoResearch Laboratories, Inc.	Cat# 712-605-150
Chemicals, Peptides, and Recombinant Proteins		
Papain	Millipore Sigma	Cat# P4762-25MG
Collagenase type I	Millipore Sigma	Cat# SCR103
E-64	Millipore Sigma	Cat# E3132-1MG
Fetal Bovine Serum	Millipore Sigma	Cat# F0926-50ML
Schneider’s Media	Millipore Sigma	Cat# S0146-500ML
DRAQ5	abcam	Cat# ab108410
Dextran sulfate, 50% solution	Millipore Sigma	Cat# S4031
Critical Commercial Assays		
10X chromium v3 single-cell gene expression kit	10X Genomics	Cat# 1000154
10X chromium v2 single-cell gene expression kit	10X Genomics	Cat# 120234
Deposited Data		
Raw reads and analyzed counts matrices	This study	GEO: GSE153723

Experimental Models: Organisms/Strains		
D. melanogaster, R9D11-Gal4 driver line: w[1118]; P{y[+t7.7] w[+mC]=GMR9D11-GAL4}attP2	BDSC	RRID:BDSC_40731
D. melanogaster, R9D11-CD4::tdTomato membrane reporter line: w[1118]; P{y[+t7.7] w[+mC]=R9D11-CD4-tdTom}attP2/TM6B, Tb[1]	BDSC	RRID:BDSC_40731
D. melanogaster: yw;;UAS-hH2B::2xmNG	This study	NA
D. melanogaster: yw;;UAS-hH2B::2xTagBFP2	This study	NA
D. melanogaster, Sp1::EGFP protein fusion reporter line: w[1118]; PBac{y[+mDint2] w[+mC]=Sp1-EGFP.S}VK00033	BDSC	RRID:BDSC_38669
D. melanogaster, UAS-IVS-myr::tdTomato membrane reporter line: w[*]; P{y[+t7.7] w[+mC]=10XUAS-IVS-myr::tdTomato}attP40	BDSC	RRID:BDSC_32222
Oligonucleotides		
mNeonGreen HCR probe set	Molecular Instruments	PRC014
CycE HCR probe set	Molecular Instruments	PRD167
D HCR probe set	Molecular Instruments	PRC881
Sp1 HCR probe set	Molecular Instruments	PRC883
TfAP-2 HCR probe set	Molecular Instruments	PRD168
Fas3 HCR probe set	Molecular Instruments	PRC900
ytr HCR probe set	Molecular	PRE680

	Instruments	
E(spl)m6-BFM HCR probe set	Molecular Instruments	PRE684
tap HCR probe set	Molecular Instruments	PRE682
jim HCR probe set	Molecular Instruments	PRE686
lncRNA:cherub HCR probe set	Molecular Instruments	PRG382
dati HCR probe set	Molecular Instruments	PRG385
mamo HCR probe set	Molecular Instruments	PRG383
bi HCR probe set	Molecular Instruments	PRG384
Software and Algorithms		
Fiji/ImageJ	Schindelin et al., 2012	RRID:SCR_002285 https://fiji.sc/
scanpy scRNA-seq analysis software	Wolf, F., Angerer, P. & Theis, F., 2018	RRID:SCR_018139
STAR RNA-seq aligner	Dobin et al., 2013	RRID:SCR_015899 https://github.com/alexdobin/STAR
Palantir pseudotime trajectory fitting software	Setty, M., Kiseliovas, V., Levine, J. et al., 2019	https://github.com/dpeerlab/Palantir
PyGAM model fitting software	Servén D, Brummitt C, Abedi H, 2018	https://github.com/dswah/pyGAM
MiCV web tool	This work	https://micv.works https://github.com/Cai-Lab-at-University-of-Michigan/MiCV

A3: Single-Cell Dissociation Protocol for Larval *Drosophila* Brains

Materials

On ice:

- 20mg/mL papain (solution in water)
- 20mg/mL collagenase
- 100X E-64 protease inhibitor
- 1X Rinaldini's solution
- Schneider's insect culture media + 10% FBS
- DRAQ5 or other DNA/live/dead-cell stain(s)

At room temperature:

- 15uM ZnCl solution (alternatively, 100mM CaCl2 solution: both are 100X working concentration for chemical digestion)

Other:

- DNA low-binding 1.5mL tubes
- Silicanized p200 pipette tips (green, brains stick less to them)
- Heat block set to 37C
- 96-well plate, non-sterile is fine (for visualizing cells post-sort)

Notes

- Keep tubes with brains on ice unless otherwise noted

Dissection

1. Prepare a 1.5mL DNA low-binding tube with 30uL 1X Rinaldini's solution
2. Dissect larval brains in cold 1X Rinaldini's solution
3. Transfer brains to 1.5mL tube from step (1)
4. Dissect for no more than 1hr total to avoid early-dissected brains from deteriorating too much relative to late-dissected brains

Chemical digestion

1. Bring the volume of the tube with dissected brains up to 80uL with 1X Rinaldini's solution (either wash and replace, or add on top of media in tube, depending on how clean your dissections are/careful you'd like to be)
2. Add 10uL of 20mg/mL papain (final conc. 2mg/mL)
3. Add 10uL of 20mg/mL collagenase (final conc. 2mg/mL)
4. Add 1uL of 15uM ZnCl solution (final conc. 150nM; Zn or Ca ions necessary for optimal collagenase activity, according to spec sheets)
5. Mix gently by flicking the tube, until you see collagenase mix uniformly and/or brains from bottom of tube come up into the solution
6. Incubate at 37C for 60min on the heat block
    - At 15min intervals, mix gently by taking the tube off of the heat block and flicking, as in step (5).
7. After incubation, place tube on ice and add 1uL E-64 100X solution; mix gently by flicking – E-64 binds papain irreversibly, quenching the digestion reaction
8. Incubate on ice for 2min
9. Spin brains down at 500G for 3min
10. Remove most of the solution using silicanized p200 tip
11. Re-suspend brains by adding Schneider's + 10% FBS solution to approximately 100uL final volume

Mechanical disruption

1. Set the volume of a p100 pipettor to 70% (approximately 70% of the total volume; adjust if using a higher volume)
2. Attach a silicanized p200 (green) pipette tip
3. Open tube, and hold up to a bright working lamp to clearly visualize brains.
4. Insert pipette tip and hold just above the bottom of the tube (about 1mm)
5. Pipette up and down in a smooth rhythm 30 times, at a rate of about 1.5 full pipetting motions every second

- Try not to be overly forceful but instead aim to maintain a consistent forward and back-pressure
- We have found this area of the protocol can be critical to dissociation success, particularly for larval VNCs where the cells seem to be more tightly coupled together – practice practice practice!

Sorting

1. Estimate cell concentration using a hemocytometer
2. Dilute dissociated cells in more Schneider's + 10% FBS to approach a desired cell concentration
   - For most commercial FACS systems, aim for at least 500uL to avoid cell loss due to 'dead volume' (volume of solution at the bottom of the tube that the FACS sample injection system cannot physically reach)
3. Transfer cells to a FACS tube appropriate for your FACS machine
4. Add cell stain(s) to appropriate tubes
   - if DRAQ5 (recommended), add 1uL of DRAQ5 stock to 500uL of dissociated cells, and do not wash afterwards
5. Sort into 1.5mL tubes pre-filled with at least 200uL of Schneider's + 10% FBS, or other capture media/vessels
6. Spin down cells after sorting at 300G for 5 min
7. Remove excess media and re-suspend at desired cell concentration
   - Assume "true" cell count in tube is ~70% of what the FACS machine reports were sorted into the tube
8. Estimate cell concentration by taking 5uL of cells and transferring to a single well of a 96-well plate, pre-loaded with 25uL of media, spinning down at 400G for 2 min, and counting FP+ cells on an epifluorescent scope
   - Hemocytometer is not reliable at low cell concentration, and may require more cells than you have. If you want, you can skip this step and proceed straight to 10X/other scRNA-seq prep platform, but the technician's might not like it!

A4: HCRv3 Staining Protocol for Larval *Drosophila* Brains

Tissue fixation and preparation

- Fix brains in 4% PFA for 20min @ RT, nutating
- Wash and permeablize brains in 0.3% PBSTw
    - 2x for 30s @ RT, standing
    - 1x for 20min @ RT, nutating
    - 1x again @ RT, nutating OR standing @ 4C ← stopping point

HCR probe hybridization

- Incubate brains in 500uL hybridization buffer (see below) @ 37C for 60min, nutating
- Add 2.5uL of 1uM probe stock for each probe (5nM final concentration)
- Incubate @ 37C overnight, nutating
- Wash 2x in wash buffer (see below) @ RT for 30min, nutating

HCR imager amplification

- Incubate brains in 500uL amplification buffer (see below) @ RT for 30min, nutating
- Snap-cool imager hairpins:
    - Each HCR probe has 2 imager hairpins that go along with it, labeled B(n)H(1/2)
    - Add 10uL of each imager hairpin stock solution (3uM) to separate PCR tubes
    - Incubate imager hairpins @ 95C for 90s, then cool to RT/4C immediately
- Add 10uL of each imager hairpin to brains (60nM final concentration for each hairpin)

- Incubate @ RT overnight, nutating
- Wash 2x with 2XSSCTw @ RT for 30min, nutating
- Wash 1x with 2XSSC @ RT indefinitely, nutating ← stopping point

Table 2.3: HCR Buffer recipes

0.3% PBSTw (approximate)

50% Tween 20	1 drop (~50uL)
1X PBS	8283uL

2X SSCTw (0.5% Tween 20, approximate)

50% Tween 20	1 drop (~50uL)
20X SSC	1000uL
Water	9000uL

Hybridization buffer

50% dextran sulfate	200uL
20X SSC	100uL
0.5% Tween 20	500uL
100% formamide	100uL
Water	100uL
Total	1000uL

Wash buffer

20X SSC	100uL
0.5% Tween 20	500uL
100% formamide	300uL
Water	100uL
Total	1000uL

Amplification buffer

50% dextran sulfate	200uL
0.5% Tween 20	500uL
20X SSC	100uL
Water	200uL
Total	1000uL

A5: Open Source Software Packages Used in This Work

Alex Wolf, Philipp Angerer, Fidel Ramirez, Isaac Virshup, Sergei Rybakov, Gokcen Eraslan, Tom White, Malte Luecken, Davide Cittaro, Tobias Callies, Marius Lange, Andrés R. Muñoz-Rojas. (2021, February 24). scanpy. Retrieved May 18, 2021, from http://github.com/theislab/scanpy (Original work published July 24, 2017)

Ali-Akber Saifee. (2020, August 25). flask-limiter. Retrieved May 18, 2021, from https://flask-limiter.readthedocs.org (Original work published February 12, 2014)

Alistair Miles. (2021, January 26). numcodecs. Retrieved May 18, 2021, from https://github.com/zarr-developers/numcodecs (Original work published September 19, 2016)

Alistair Miles. (2021, April 16). zarr. Retrieved May 18, 2021, from https://github.com/zarr-developers/zarr-python (Original work published December 18, 2015)

Anaconda, Inc. (2020, June 30). numba. Retrieved May 18, 2021, from http://numba.github.com (Original work published August 15, 2012)

Andy McCurdy. (2020, June 1). redis. Retrieved May 18, 2021, from https://github.com/andymccurdy/redis-py (Original work published October 8, 2012)

Armin Ronacher. (2020, April 3). flask. Retrieved May 18, 2021, from https://palletsprojects.com/p/flask/ (Original work published April 16, 2010)

Armin Ronacher. (2021, March 18). flask-sqlalchemy. Retrieved May 18, 2021, from https://github.com/pallets/flask-sqlalchemy (Original work published June 2, 2010)

Ask Solem. (2020, December 16). celery. Retrieved May 18, 2021, from http://celeryproject.org (Original work published April 27, 2009)

Benedikt Schmitt. (2019, May 18). filelock. Retrieved May 18, 2021, from

https://github.com/benediktschmitt/py-filelock (Original work published July 6, 2014)

Benoit Chesneau. (2021, April 27). gunicorn. Retrieved May 18, 2021, from https://gunicorn.org (Original work published January 3, 2010)

Chris Parmer. (2021, April 8). dash. Retrieved May 18, 2021, from https://plotly.com/dash (Original work published June 20, 2017)

Cloudpipe. (2020, August 25). cloudpickle. Retrieved May 18, 2021, from https://github.com/cloudpipe/cloudpickle (Original work published April 16, 2015)

Daniel Serven. (2018, October 31). pygam. Retrieved May 18, 2021, from https://github.com/dswah/pyGAM (Original work published May 10, 2017)

Denis Bilenko. (2021, January 20). gevent. Retrieved May 18, 2021, from http://www.gevent.org/ (Original work published July 20, 2009)

Elmer Thomas, Yamil Asusta. (2021, April 21). sendgrid. Retrieved May 18, 2021, from https://github.com/sendgrid/sendgrid-python/ (Original work published June 28, 2012)

Faculty. (2021, March 21). dash-bootstrap-components. Retrieved May 18, 2021, from https://dash-bootstrap-components.opensource.faculty.ai/ (Original work published September 21, 2018)

Flask-Admin team. (2021, April 17). flask-admin. Retrieved May 18, 2021, from https://github.com/flask-admin/flask-admin/ (Original work published July 11, 2011)

Francesc Alted, Valentin Haenel. (2020, September 9). blosc. Retrieved May 18, 2021, from http://github.com/blosc/python-blosc (Original work published November 17, 2010)

Golovanov Stanislav. (2017, January 9). pdfkit. Retrieved May 18, 2021, from UNKNOWN (Original work published January 8, 2013)

Grey Li. (2021, May 18). bootstrap-flask. Retrieved May 18, 2021, from https://github.com/greyli/bootstrap-flask (Original work published June 11, 2018)

Inada Naoki. (2020, December 18). msgpack. Retrieved May 18, 2021, from https://msgpack.org/ (Original work published January 6, 2018)

Inkscape Project. (2020). *Inkscape*. Retrieved from https://inkscape.org

Jean-Christophe Fillion-Robin. (2020, November 11). cmake. Retrieved May 18, 2021, from http://cmake.org/ (Original work published November 9, 2016)

Jonathan Underwood. (2020, November 18). lz4. Retrieved May 18, 2021, from https://github.com/python-lz4/python-lz4 (Original work published January 31, 2012)

Joshua Tauberer. (2020, November 5). email_validator. Retrieved May 18, 2021, from https://github.com/JoshData/python-email-validator (Original work published April 21, 2015)

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., … Willing, C. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90).

Konsta Vesterinen. (2020, June 2). wtforms_alchemy. Retrieved May 18, 2021, from https://github.com/kvesteri/wtforms-alchemy (Original work published January 26, 2013)

Krzysztof Polanski, Jongeun Park. (2020, June 9). bbknn. Retrieved May 18, 2021, from https://github.com/Teichlab/bbknn (Original work published July 19, 2018)

Leland McInnes. (2020, May 15). umap-learn. Retrieved May 18, 2021, from http://github.com/lmcinnes/umap (Original work published November 10, 2017)

Leonard Richardson. (2020, October 3). beautifulsoup4. Retrieved May 18, 2021, from http://www.crummy.com/software/BeautifulSoup/bs4/ (Original work published October 2, 2013)

magic-impute. (2019, November 18). Retrieved May 18, 2021, from https://github.com/KrishnaswamyLab/MAGIC (Original work published July 24, 2018)

Manu Setty. (2020, May 20). palantir. Retrieved May 18, 2021, from https://github.com/dpeerlab/palantir (Original work published March 6, 2020)

Matt Wright & Chris Wagner. (2020, July 28). flask-security-too. Retrieved May 18, 2021, from https://github.com/Flask-Middleware/flask-security (Original work published April 25, 2019)

Niko Pasanen <niko@pasanen.me>. (2020, October 27). dash-uploader. Retrieved May

18, 2021, from https://github.com/np-8/dash-uploader (Original work published April 26, 2020)

Patrick Vogel, Bogdan Petre. (2020, September 9). flask_monitoringdashboard. Retrieved May 18, 2021, from https://github.com/flask-dashboard/Flask-MonitoringDashboard (Original work published February 27, 2018)

Peter Justin. (2020, June 2). flask_caching. Retrieved May 18, 2021, from https://github.com/sh4nks/flask-caching (Original work published July 4, 2016)

Philipp Angerer, Alex Wolf, Isaac Virshup, Sergei Rybakov. (2021, April 11). anndata. Retrieved May 18, 2021, from http://github.com/theislab/anndata (Original work published November 7, 2017)

plotly. (2018, December 19). dash-dangerously-set-inner-html. Retrieved May 18, 2021, from https://pypi.org/project/dash-dangerously-set-inner-html/ (Original work published December 7, 2017)

Seth M. Morton. (2020, November 21). natsort. Retrieved May 18, 2021, from https://github.com/SethMMorton/natsort (Original work published November 17, 2012)

Stochastic Technologies. (2020, March 6). shortuuid. Retrieved May 18, 2021, from https://github.com/stochastic-technologies/shortuuid/ (Original work published January 8, 2011)

The Biopython Contributors. (2020, May 25). biopython. Retrieved May 18, 2021, from https://biopython.org/ (Original work published April 28, 2009)

The Python Cryptographic Authority developers. (2020, August 16). bcrypt. Retrieved May 18, 2021, from https://github.com/pyca/bcrypt/ (Original work published May 11, 2013)

Tom McCarthy. (2020, June 21). checksumdir. Retrieved May 18, 2021, from http://github.com/cakepietoast/checksumdir (Original work published March 4, 2015)

(Unknown). (2020, September 24). toolz. Retrieved May 18, 2021, from https://github.com/pytoolz/toolz/ (Original work published September 11, 2013)

V.A. Traag. (2020, September 23). leidenalg. Retrieved May 18, 2021, from https://github.com/vtraag/leidenalg (Original work published October 23, 2018)

Valentin LAB. (2017, November 19). colour. Retrieved May 18, 2021, from http://github.com/vaab/colour (Original work published April 14, 2013)

Willighagen LG. 2019. Citation.js: a format-independent, modular bibliography tool for the browser and command line. PeerJ Computer Science 5:e214 https://doi.org/10.7717/peerj-cs.214

Yiming Yang, Bo Li. (2020, July 26). harmony-pytorch. Retrieved May 18, 2021, from https://github.com/lilab-bcb/harmony-pytorch (Original work published January 16, 2020)